

Pocket Evidence Based Medicine

A Survival Guide for Clinicians
and Students

Walter R. Palmas

 Springer

Pocket Evidence Based Medicine

Walter R. Palmas

Pocket Evidence Based Medicine

A Survival Guide for Clinicians
and Students

 Springer

Walter R. Palmas
New York, NY, USA

ISBN 978-3-031-19470-2 ISBN 978-3-031-19471-9 (eBook)
<https://doi.org/10.1007/978-3-031-19471-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

“That’s great. But, isn’t everything you do in Medicine already evidence-based?” a dear friend pointedly asked after I told him I taught Evidence-Based Medicine. His question reflected the natural expectation our patients have about all our clinical decisions—that they are based on studies, and that we are familiar with the available data and confident in our assessment of new publications. Asking clinically relevant questions and evaluating the literature in search for answers is an enriching and essential component of any career in healthcare. However, the massive breadth and depth of medical publications pose a substantial challenge to students and healthcare professionals alike.

This brief book has been written as a tool to guide the reader of medical research, taking a practical, concise, and clinically oriented approach. It is based on what I learned over many years of teaching Evidence-Based Medicine to Medicine residents and medical students at Columbia University Medical Center. Teaching smart young people has been the greatest privilege of my medical career, and I have always viewed it as a modest attempt to pay back for the generous and masterful teachings I had received.

Early in my career, I was extremely fortunate to learn from George A. Diamond MD, at Cedars-Sinai Medical Center in California. George was a brilliant thinker, and he pioneered the innovative application of Bayesian methods and a rigorous evidence-based approach to Cardiology. He was also an extremely patient teacher and had a wicked sense of humor. I hope my work, including this book, does a modicum of justice to his mentorship.

New York, NY, USA

Walter R. Palmas

Contents

1	The Most Basic Concepts in Biostatistics	1
1.1	Statistical Inference: From a Sample to the Population	1
1.2	Internal and External Validity	1
1.3	Null Hypothesis Significance Testing	2
1.4	The Almighty P -Value	2
1.5	Limitations of P -Values	3
1.6	More About the Meaning of the P -Value	4
1.7	Two-Sided P -Values Are the Norm in Medicine	4
1.8	Confidence Intervals	5
1.9	Type 1 and Type 2 Statistical Errors	6
1.10	We All Do Frequentist Statistics	8
1.11	Bayesian Statistics: Credible Intervals, Probability Estimates	9
1.12	Confounding	10
1.13	Interaction or Effect Modification	13
1.14	Collider Bias	14
2	Assessment of Diagnostic Tests	15
2.1	Sensitivity, Specificity, Predictive Values	15
2.2	Likelihood Ratios	17
2.3	Receiver Operating Characteristic (ROC) Curves	17
2.3.1	Effect of Threshold Changes on Predictive Values	22

2.4	F-Score	23
2.5	Common Biases in the Evaluation of Diagnostic Tests	24
2.5.1	Spectrum Bias	24
2.5.2	Post-Test Referral Bias	25
2.5.3	Biased Gold Standard	28
2.5.4	Highly Selected Populations	29
2.6	Avoiding Biases	30
2.7	Checklist for the Assessment of a Diagnostic Test	31
2.8	Screening	32
3	Use of a Diagnostic Test	35
3.1	The Two-by-Two Table in Different Scenarios	35
3.2	Pretest Probability Estimates	38
3.3	Likelihood Ratios and Fagan Nomogram	38
3.4	Use of Predictive Models	40
4	Observational Studies	43
4.1	Observational Studies. General Considerations	43
4.2	Cross-Sectional Studies	45
4.3	Odds Ratio	46
4.4	Case Control Studies	46
4.4.1	Importance of the Control Group	47
4.4.2	Ascertainment or Diagnostic Bias	48
4.4.3	Recall Bias	48
4.4.4	Interviewer Bias	49
4.4.5	Nested Case-Control Studies	49
4.5	Prospective Cohort Studies	50
4.5.1	Selection Bias	51
4.5.2	Confounding by Indication (Prescription Bias)	52
4.5.3	Immortal Time Bias	56
4.5.4	Attrition of those Susceptible	57
4.5.5	Protopathic Bias	59

4.5.6	Chronology (Secular) Bias.	60
4.5.7	Non-Randomized Outcomes Studies.	60
4.5.8	Checklist for Observational Studies.	61
5	Commonly Used Statistics	63
5.1	Relative Risk	63
5.2	Relative Risk Reduction.	65
5.3	Number Needed to Treat	65
5.4	Number Needed to Harm.	67
5.5	Censoring.	67
5.6	Kaplan-Meier Curves.	68
5.6.1	Incorrect Kaplan-Meier Formatting.	71
5.6.2	Censoring and Numbers at Risk	72
5.7	Hazard Ratio	74
5.7.1	Benefits of an Adjusted Hazard Ratio	75
5.7.2	Assessing Palliative Treatments.	76
5.7.3	The Proportionality Assumption	78
5.8	The Log-Rank Statistic	79
5.8.1	Observed Vs. Expected Event Rates in Month 1	80
5.8.2	Observed Vs. Expected Event Rates in Month 2	80
5.8.3	Observed Vs. Expected Event Rates in Month 3	81
5.8.4	Obtain a Global Chi-Square Test.	82
5.9	Odds Ratio	83
5.10	Vaccine Efficacy.	84
5.11	Attributable Proportion	85
6	Randomized Clinical Trials	87
6.1	Why Do We Need Randomized Trials?	87
6.2	Types of Clinical Trials by Phase.	88
6.3	Clinical Trial Registration and Compliance with Guidelines	88
6.4	Inclusion and Exclusion Criteria	89
6.5	Internal and External Validity	90
6.6	Type 1 and Type 2 Statistical Errors	90

6.7	Sample Size and Power Estimates	91
6.8	Randomization: The R in RCT	92
6.8.1	Preventing Bias	92
6.8.2	Reducing Confounding	93
6.8.3	Stratified Randomization	94
6.8.4	Randomization by Blocks	94
6.8.5	Did Randomization Prevent Confounding?	96
6.8.6	Clustered Randomization	96
6.8.7	Fixed Unequal Allocation	98
6.8.8	Dynamic Allocation: Adaptive Randomization and Minimization	98
6.9	Blinding	100
6.9.1	Involuntary Unmasking	101
6.10	Crossovers	102
6.11	Completeness of Follow-Up	102
6.12	Intention to Treat Analysis	103
6.13	Sequential Stopping Boundaries	104
6.14	Improvements in Trial Monitoring	108
6.15	Primary and Secondary Outcomes	109
6.16	Hierarchical Testing of Secondary Outcomes	110
6.16.1	Stepwise Hierarchical Testing	111
6.17	Subgroup Analysis	112
6.18	Adaptive Clinical Trials	115
6.19	Checklist for a Randomized Clinical Trial	117
6.20	Assessing a Negative Clinical Trial	118
6.21	Checklist for a Negative Clinical Trial	119
7	Non-inferiority Clinical Trials	121
7.1	Why Do We Need Non-Inferiority Trials?	121
7.2	A Quick Reminder about Superiority Trials	121
7.3	Hypotheses in Superiority Trials	122
7.4	Type 2 Error in Superiority Trials	123
7.5	A Whole New World: Hypotheses in Non-Inferiority Trials	123

7.6	The Relative Risk (RR) Non-Inferiority Margin	125
7.7	The Absolute Risk Difference (ARD) Non-Inferiority Margin	127
7.8	Both Relative and Absolute Risk Are Clinically Important	128
7.9	Type 1 Error in Non-Inferiority Trials	128
7.10	As Treated and Intention to Treat Analysis	129
7.11	Superiority Analysis in a Non-Inferiority Trial	130
7.12	ABSORB III: A Non-Inferiority Study that Used an Absolute Risk Difference Margin	130
7.13	Is a Non-Inferiority Trial Acceptable when the Outcome Is Death?	132
7.14	Checklist for Non-Inferiority Trials	133
8	Bayesian Analysis of Clinical Trials	135
8.1	Limitations of Conventional Statistics	135
8.2	Similarities with the Diagnostic Process	136
8.3	Prior Probability, Data, and Posterior Probability	137
8.4	Types of Prior Probability	137
8.5	Prior Probabilities and Clinical Common Sense	138
8.6	The Use of Excessively Optimistic Priors Should be Avoided	139
8.7	Advantages of the Uninformative Prior	141
8.8	Navigating Statistical Lingo in the Methods Section	142
8.9	Clinical Interpretation of Posterior Probabilities	142
8.10	Irrelevance of the Null Hypothesis	144
8.11	The Added Value of Bayesian Methods	144
9	Health Economics Studies	145
9.1	Basic Definitions	145
9.2	Commonly Used Outcome Measurements	146

9.3	Controversial Issues	147
9.4	Sponsorship Bias	148
9.5	Role of Health Economics Studies in Healthcare Policy Making	149
9.6	Checklist for Health Economics Studies	150
10	Meta-Analysis	151
10.1	Systematic Reviews and Meta-Analysis	151
10.2	Publication Search	152
10.3	Publication Bias	153
	10.3.1 Funnel Plot.	153
	10.3.2 Significance Tests	155
10.4	Selective Reporting	156
10.5	Quality of Individual Studies—Risk of Bias	158
10.6	Testing for Heterogeneity	159
	10.6.1 How Much Do Individual Study Results Differ from each Other?	159
10.7	Obtaining a Pooled Estimate	161
	10.7.1 Fixed Effect Model	161
	10.7.2 Random Effects Model	162
	10.7.3 Fixed Versus Random Effects	162
10.8	Sensitivity Analysis	164
10.9	Meta-Regression	165
10.10	Network Meta-Analysis	165
	10.10.1 Qualitative Assessment of Transitivity.	166
	10.10.2 Quantitative Assessment of Consistency (a.K.a. Coherence)	166
10.11	Meta-Analysis Checklist	170
11	Introduction to Artificial Intelligence	
	Methods	173
11.1	Basic Definitions	173
11.2	Performance Analysis of Machine Learning Models	175
11.3	Biases in Artificial Intelligence	176

11.4	Quality of Medical Research Using Artificial Intelligence	177
11.5	Checklist for Artificial Intelligence Methods	178
12	Finding the Best Evidence	181
12.1	The Essential Components of Clinical Practice.	181
12.2	Five Step EBM Model	182
12.3	The Hierarchy of Evidence	182
12.4	The Cochrane Collaboration	183
12.5	PubMed Queries.	183
12.6	Other Reliable Sources	184
12.7	Google Scholar.	184
12.8	The Future of EBM	185
13	Ethics of Clinical Research.	187
13.1	Relevance.	187
13.2	The Belmont Report.	188
13.3	Beneficence	188
13.4	Justice.	189
13.5	Respect for Autonomy	189
13.6	Institutional Review Boards.	189
13.7	Conflicts of Interest	190
	Appendix: Self-Assessment Test	193
	Answers to Practice Questions.	215
	Open Access Evidence Based Clinical Knowledge.	217
	Online EBM Resources That Require Institutional Access	219
	EBM Calculators, Decision Making Tools	221



The Most Basic Concepts in Biostatistics

1

1.1 Statistical Inference: From a Sample to the Population

Statistics encompasses the methods of collecting, summarizing, analyzing, and drawing conclusions from data.

Biostatistics, in turn, is the application of statistics to biological, medical, and public health data.

In Medicine we are interested in preventing, diagnosing, and treating diseases in big Populations. But those Populations are too large to be studied in their entirety. For that reason, we obtain what we believe is a Representative Sample from that Population for our study.

A certain characteristic of a Population is known as a Parameter, whereas that characteristic in a Sample is known as a Variable.

Statistical Inference is the process of drawing conclusions about Parameters in a Population from the assessment of Variables in a Sample.

1.2 Internal and External Validity

Internal validity is the robustness of our experiment, including our design and sample, to optimally test our hypothesis. In other

words, it is the ability to ensure that a true association or effect will be appropriately captured by the experiment.

External validity is the ability to provide results from our sample that are reproducible and valid in the population of interest.

1.3 Null Hypothesis Significance Testing

In Medicine a Hypothesis is a yet unproven idea about what causes a disease, how to diagnose it, or how to treat it.

This may seem strange to you, but we do not actually test the Hypothesis when we analyze the data from a study. Instead, we define and test our study data against its logical opposite, the Null Hypothesis.

The significance test we apply to the study data evaluates that data against the Null Hypothesis. We test how significantly different from the Null are data is. Therefore, the entire approach is known as Null Hypothesis Significance Testing (NHST, if you need more acronyms in your life).

1.4 The Almighty P -Value

The main product of Null Hypothesis Significance Testing is the P -Value.

The P -Value is the probability of observing an effect or difference as large as, or larger than ours if the null hypothesis were true.

The P -Value represents the strength of the evidence against the null hypothesis. Smaller P values indicate stronger evidence against the null hypothesis, and if the P -Value is small enough, we feel confident rejecting the null hypothesis.

Ronald A. Fisher proposed back in the 1920's that a probability level $< 5\%$ (i.e. a *P*-Value < 0.05) could be a useful benchmark for concluding that fairly strong evidence exists against the Null Hypothesis. His landmark book *Statistical Methods for Research Workers* reinforced that position, and provided tables for commonly used statistics with the corresponding *P*-Values. Such tables were the computational tool for most statistical work at that time, and cemented the approach.

However, it should be kept in mind that the $p < 0.05$ was never intended as an absolute threshold--the strength of evidence is on a continuum. In addition, statistics is not a religion, and the scientific and societal contexts are essential in examining any statistical result, including *P*-Values.

1.5 Limitations of *P*-Values

***P*-Values give no substantive information about the clinical importance of the result. A very large study may produce a “very significant” *P*-Value based on a small effect, which may not be relevant when translated into clinical practice.**

On the other hand, a study without enough statistical power may be indicative of a clinically important effect, even though it does not reach “statistical significance”.

Remember to always examine the effect size and confidence intervals and consider the clinical and societal context, and all other pertinent information, such as a Bayesian analysis of the data, and meta-analysis with similar studies if appropriate.

1.6 More About the Meaning of the *P*-Value

In 2016 the American Statistical Association reminded us of the following:

1. *P*-Values also indicate how incompatible the data are with a specified statistical model. In other words, a non-significant *P*-Value may also be obtained because our statistical model is not appropriate for the data at hand.
2. *P*-Values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone. We wish they did, but they don't.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *P*-Value passes a specific threshold.
4. Proper inference requires full reporting and transparency: beware of selective reporting.
5. A *P*-Value, or statistical significance, does not measure the size of an effect or the importance of a result. For that we need effect size estimates, such as Hazard Ratio with its Confidence Interval, and Number Needed to Treat.

[Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA's Statement on *P*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129–133, <https://doi.org/10.1080/00031305.2016.1154108>].

1.7 Two-Sided *P*-Values Are the Norm in Medicine

Almost all *P*-Values in healthcare research are two-sided. They reflect the assessment of whether the new treatment is harmful or beneficial. A *P*-Value <0.05 can be seen because the treatment

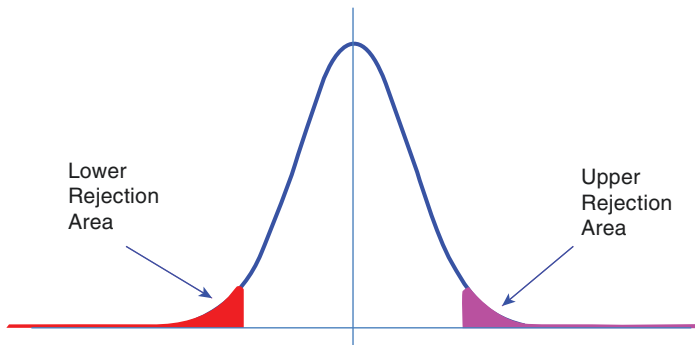


Fig. 1.1 Null Hypothesis Rejection Areas. Results falling within the shaded areas at either extreme of the distribution would make us reject the Null Hypothesis

significantly improved outcomes compared to the control arm, or because it significantly worsened outcomes.

For example, let us say we performed a study comparing the survival benefit of implanting a cardiac defibrillator in heart failure patients, as compared to usual medical care. After running a test that compares mortality between arms over follow up, a test statistic differing significantly from zero difference would allow us to reject the Null Hypothesis. However, that statistic could differ significantly from zero because it shows either a significant increase or a significant decrease in mortality. We would reject the Null Hypothesis in either case. In Fig. 1.1 the zero difference statistic is represented by the central vertical line. The two Null Hypothesis rejection areas are shaded, on both extremes.

For the same reasons, most of the confidence intervals (which we discuss next) are also two-sided.

1.8 Confidence Intervals

We use inferential statistics to obtain a Point Estimate of the population parameter, and the Confidence Interval around it. Relying on information from a sample will always lead to some level of uncertainty (the opposite of which is Confidence). A Confidence

A 95% Confidence Interval (CI) means that, if we were to repeat the same experiment numerous times, 95% of the CIs we obtain in those repeats are expected to contain the true population parameter.

Interval (CI) is thus a range of values that aims to quantify that uncertainty.

The width of the confidence interval informs our confidence on that point estimate: a wide confidence interval suggests we are not that confident in our point estimate, whereas a narrow confidence interval means we are confident about it.

The confidence interval is also informative in regards to the statistical and clinical significance of our results.

For example, if we are studying a new treatment and the 95% CI of our Relative Risk estimate does not include the Relative Risk of 1, the corresponding *P*-Value should be <0.05 . Remember, a Relative Risk of 1 means there is no difference in outcomes between the treatment groups and constitutes the Null Hypothesis.

If the 95% CI does include the Null Hypothesis value, the *P*-Value is expected to be ≥ 0.05 .

From a clinical perspective, a narrow confidence interval around a Relative Risk that shows a nice mortality reduction by a new treatment makes clinicians very happy because we are more confident about the therapeutic benefit we have detected. On the other hand, we would not be confident if the interval was very wide.

1.9 Type 1 and Type 2 Statistical Errors

When dealing with the two types of Statistical Error you do not need to memorize which one is which. Just remember human nature and it will be easy to figure out which one is Type 1.

Nobody studies a new drug, or tests a new surgical procedure with the intention to prove that it does not work—to the contrary, they hypothesize that it works, and they would love to reject the

Null Hypothesis. In other words, researchers are naturally biased towards having a positive finding. Therefore, we must be very cautious to prevent False Positive studies, in which the Null Hypothesis is wrongfully rejected. We call that wrongful rejection

Type 1 Statistical Error is the Probability of a False Positive Finding.

of the Null Hypothesis a Type 1 Statistical Error.

We take Type 1 error very seriously, and we thus set a stringent significance threshold when we assess *P*-Values. We usually require the probability of a false positive finding to be <5%.

In turn, Type 2 Error is the failure to reject the Null Hypothesis

Type 2 Statistical Error is the Probability of a False Negative Finding.

when it is actually false.

The alternative to Type 2 Error is our Statistical Power. For that reason, we estimate Power as follows:

$$\text{Power} = (1 - \text{Probability of Type 2 Error})$$

We usually require that the power of a study be 80% or more. We want to keep Type 2 error probability at 20% or less. We will later review in detail what determines the power of Randomized Clinical Trials, but let us start by stating the obvious: the larger your sample size is, the greater your statistical power shall be.

You can appreciate that we tolerate more type 2 Error (up to 20%) than Type 1 (less than 5%). From a historical perspective it has been an excellent idea to be tougher on type 1 error. Once a “positive” study is published there is little enthusiasm to repeat it, and thus Type 1 error may go unchallenged in the literature. That

is particularly true when the false positive finding came from an expensive study funded by a private party, whose agenda was fulfilled by the false positive result. On the other hand, if you have several small but otherwise well designed studies that did not individually reach statistical significance, a meta-analysis of those studies can address the issue of type 2 error, “adding up” their results. In other words, Type 2 Error may be corrected through meta-analysis, whereas we may be stuck forever with a false positive finding.

1.10 We All Do Frequentist Statistics

The definitions of P -Values and CIs we just reviewed imply a large number of theoretical repeats of the same experiment. Repeats that nobody really expects will ever take place. This approach is part of what has been the dominant school of Statistics since the early twentieth Century, and it made a whole lot of sense when the same experiments were carried out, over and over again, using beans, seeds, or plants. Because it relies on the expected frequency of results in numerous repeats of the experiment, this school is known as Frequentist Statistics. It is what we all do when we run a statistical test and obtain a P -Value.

However, Frequentist Statistics in spite of their predominance do not estimate the probability that your findings represent the true population value. For example, frequentist statistics cannot tell you anything about the probability that a new treatment reduces the risk of death by a certain amount. Or the probability that exposure to a toxic substance will cause serious harm. You need Bayesian statistics if you are interested in estimating Probabilities like those.

1.11 Bayesian Statistics: Credible Intervals, Probability Estimates

All clinicians are familiar with the Bayesian approach to diagnosis. Before obtaining a diagnostic test they estimate the pre-test probability of the disease, or prior probability of disease. After performing the diagnostic test, they integrate the test results with

Bayesian methods combine the Prior Probability with our Data to give us a Posterior Probability of the Estimate.

the prior probability estimate, and thus obtain a posterior probability of the disease. This approach makes the best possible use of all the information.

Similarly, we can interpret the results of a study or experiment, considering the prior probability of expected results, and the actual study results.

Why bother with all that? Bayesian methods can provide estimates such as “The probability that the risk of death is reduced by the new treatment”, or “The probability that the risk of death is

A 95% Credible Interval describes the boundaries within which we believe there is a 95% probability the true population value lies.

increased by 10% or more after exposure to the risk factor” Those, and many other probability estimates, can be very informative when added to the frequentist *P*-Value and the Confidence Intervals.

Credible Intervals are the Bayesian equivalent of Confidence Intervals.

In other words, given the prior probability and the observed data the actual population value has a 95% probability of falling within the estimated interval. Credible Intervals are far more intuitive than Confidence Intervals. Bayesian statistics can be easily added to frequentist statistics, but they were not favored early on because they can be computationally intensive. However, you can now run Bayesian simulations online using your phone. People in Finance, Physics, and Marketing routinely use Bayesian techniques. Technological advances and a generational change will hopefully result in a much wider use of the Bayesian approach in

To be a confounder, a variable must meet two conditions:

1. **It must be associated with the exposure.**
2. **It must be associated with the outcome, independently of the exposure. In other words, it should not be a mediator between the exposure and the outcome.**

Medicine, as a useful addition to good old Frequentist Statistics.

1.12 Confounding

We must be aware of the possibility of confounding in all clinical research studies (Fig. 1.2).



Fig. 1.2 Confounding. The confounder is associated with both the exposure and the outcome, and it is not a mediator

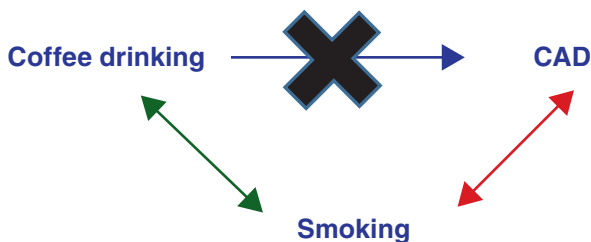


Fig. 1.3 Adjustment. Adjusting for smoking removes the perceived association between coffee and CAD

A notorious example of confounding can be found in the old studies reporting that coffee drinking (exposure) was associated with the risk of Coronary Artery Disease (outcome). Those studies had one major problem: they did not consider the fact that in those days many people who drank a lot of coffee also smoked cigarettes. Once you adjust for smoking (the confounder) there is no association between coffee drinking and CAD (Fig. 1.3).

In Randomized Clinical Trials (RCT) of relatively small size we can have trouble with confounding if we fail to stratify our randomization by the potential confounders. Let us say we compare the effect of atorvastatin versus placebo to reduce the risk of incident Myocardial Infarction (MI) in a small RCT. If we do not stratify our randomization by smoking status we may be unlucky and end up with more smokers in one arm of the study, as shown in Fig. 1.4.

When more smokers were randomized to receive atorvastatin, a spurious association of smoking with atorvastatin was created.

This would be problematic, because smoking may decrease the observable benefit of atorvastatin, even if we adjust our analysis to address the confounding, after the study is completed.

In smaller RCT's we may address the risk of potential confounding using one of the following methods:

- (a) Excluding people with a strong confounding characteristic (e.g., we may exclude all people with advanced Chronic

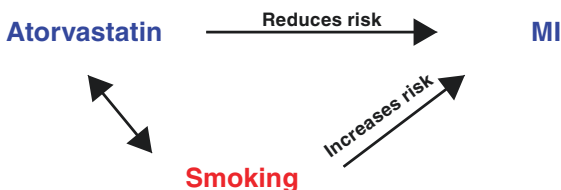


Fig. 1.4 Confounding Caused by Inappropriate Randomization. More smokers were assigned to the atorvastatin arm, creating a spurious association with the cholesterol lowering drug and thus causing confounding

Kidney Disease). This would reduce the pool of potential participants and the applicability of our findings.

An Interaction is present when the third variable (the effect modifier) changes significantly the relationship between exposure and outcome, without being associated to either of them.

The association between the exposure and the outcome is different for each level of the effect modifier.

- (b) Stratifying our randomization within strata of the confounder (i.e., randomize smokers separately from non-smokers) to ensure a balanced distribution of that characteristic across study arms. This is the best approach.
- (c) Adjusting for confounders in the analysis.



Fig. 1.5 Contraceptives and VTE Risk. Contraceptive use causes a very small increment in VTE risk

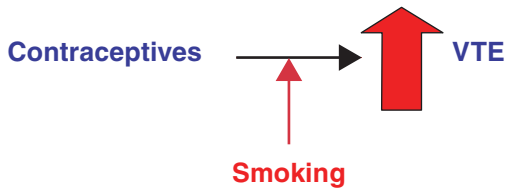


Fig. 1.6 Effect Modification. Smoking greatly increases the risk of VTE in contraceptive users

1.13 Interaction or Effect Modification

For example, when we prescribe an oral contraceptive to a healthy young lady who is not severely obese the risk of provoking a Venous Thromboembolism (VTE) is almost negligible, as depicted below (Fig. 1.5).

It is negligible, indeed, unless the patient smokes. Smoking substantially increases the risk of VTE (Fig. 1.6). Smoking is an effect modifier of the association of interest.

Obviously, taking a contraceptive does not make you crave tobacco, and smoking by itself does not cause VTE. In other words, smoking, the effect modifier, is not associated with either the exposure or the outcome but fundamentally changes the association between the exposure and the outcome.

In observational studies we deal with Interaction/Effect Modification as follows:

1. We test the significance of an interaction term in multivariate models.
2. If the interaction is significant, we perform a stratified analysis, reporting the strength of the association between exposure and outcome for each stratum of the Effect Modifier. In our example, we should report the risk of VTE for smokers and non-smokers separately.

In interventional studies (randomized clinical trials) we explore the results for interactions in the subgroup analysis. This explora-

tion may generate a hypothesis for further studies, but is usually not viewed as strong enough to reject additional null hypotheses.

1.14 Collider Bias

A third variable that is independently caused by both the putative exposure and the putative outcome is known as a collider. When a study controls by the collider, either through sampling or statistical analysis, collider bias is created and we erroneously characterize the association between exposure and outcome. Collider bias may cause us to believe that there is an association when there is none. The original description, like so much of our understanding of EBM, was given to us by David Sackett. He described the association detected between locomotor diseases and respiratory diseases in a sample of 257 hospitalized patients. However, when he performed a similar analysis in a sample of 2783 subjects from the general population, the association was no longer detected. Both locomotor and pulmonary disease independently cause hospitalizations—hospitalization is a collider variable because the two causal pathways run into each other at the hospital admission. Therefore, the analysis performed in hospitalized patients suffered from collider bias caused by sampling. Of note, if the analysis performed in 2783 subjects from the general population had been adjusted by a “hospitalization” variable a similar collider bias could have been created—this time through an analytic error.

In summary, it is important to remember that collider bias may be caused at the time of sampling or analysis when we fail to identify the collider variable as such.

Assessment of Diagnostic Tests

2

2.1 Sensitivity, Specificity, Predictive Values

Most of us are familiar with these definitions, so let us keep it brief, starting with a 2 x 2 table (Table 2.1).

Sensitivity is the proportion of people WITH THE DISEASE who have a POSITIVE TEST. In other words, it reflects the probability of having an abnormal test among those who have the disease.

$$\text{Sensitivity} = \text{True Positives} / (\text{True Positives} + \text{False Negatives})$$

Specificity is the proportion of people FREE OF DISEASE who have a NEGATIVE TEST. It reflects the probability of having a normal test result among those who are disease free.

$$\text{Specificity} = \text{True Negatives} / (\text{True Negatives} + \text{False Positives})$$

Table 2.1 Two-by-two diagnostic table

	Disease present	Disease absent
Abnormal test	True positive	False positive
Normal test	False negative	True negative

Positive Predictive Value (PPV) is the proportion of people WITH A POSITIVE TEST who have the DISEASE. In other words, it reflects the probability of having the disease among those who tested positive.

$$\text{PPV} = \text{True Positives} / (\text{True Positives} + \text{False Positives})$$

A higher PPV is obtained when the test Specificity and/or the Disease Prevalence are higher.

Negative Predictive Value (NPV) is the proportion of people with a NEGATIVE TEST who are DISEASE FREE. In other words, it reflects the probability of not having disease among those who tested negative.

$$\text{NPV} = \text{True Negatives} / (\text{True Negatives} + \text{False Negatives})$$

A higher NPV is obtained when the test Sensitivity is higher and/or when the disease Prevalence is lower.

The effect of Prevalence on the predictive values is worth remembering, because it is independent from the test's Sensitivity and Specificity. When reporting their findings some researchers are quite emphatic about the importance of the good predictive values of their diagnostic test. You should remember that, for the same sensitivity and specificity, high prevalence increases the PPV, and low prevalence increases the NPV. In other words, the predictive values quoted in a given sample will be reproducible in your population only if the prevalence is similar.

Finally, we often wish to assess the probability of the disease being present in those who had a normal (negative) test result. We estimate that as follows:

$$1 - \text{NPV} = \text{False Negatives} / (\text{True Negatives} + \text{False Negatives})$$

2.2 Likelihood Ratios

Likelihood is a synonym of Probability. The Likelihood Ratios are therefore ratios of two probabilities. They give us useful information because they compare the probability of a given type of test result, positive or negative, in those who have the disease, and in those who do not have the disease.

The Positive Likelihood Ratio, LR (+), is defined as the probability of a positive test among individuals with the disease, relative to the probability of the same positive test but among those without disease.

$LR (+) = \text{Prob. Positive Test in Disease} / \text{Prob. Positive Test in Disease-Free}$. We estimate it as follows:

$$\begin{aligned} LR (+) &= (TP / TP + FN) / (FP / FP + TN) \\ &= \text{Sensitivity} / (1 - \text{Specificity}) \end{aligned}$$

The Negative Likelihood Ratio, LR (−), is defined as the probability of a negative test among individuals with the disease, relative to the probability of the same negative test but among those without disease.

$LR (-) = \text{Prob. Negative Test in Disease} / \text{Prob. Negative Test in Disease-Free}$. We estimate it as follows:

$$\begin{aligned} LR (-) &= (FN / TP + FN) / (TN / FP + TN) \\ &= (1 - \text{Sensitivity}) / \text{Specificity}. \end{aligned}$$

We will discuss in detail how to use Likelihood Ratios in the Chapter about Use of Diagnostic Tests.

2.3 Receiver Operating Characteristic (ROC) Curves

When we create a diagnostic 2x2 table we consider only two possible test results: Normal or Abnormal. However, for most tests we have either different degrees of abnormality (such as High

Probability, Intermediate Probability, and Low Probability), or a continuous type of result (such as pro-BNP level or Troponin level).

In those situations, the Sensitivity and Specificity will change, in opposite directions, depending on which value we consider diagnostic for the disease. In other words, the Sensitivity and Specificity will change, in opposite directions, depending on the cutoff value that we consider abnormal for the test.

We can assess how well a test performs across all its possible values plotting the Sensitivity values that correspond to each [1-Specificity] value. That generates a Receiver Operating Characteristic (ROC) curve.

Let us see how that worked in a classic study. Maisel et al. studied 1586 patients who went to the emergency department with acute dyspnea and whose B-type natriuretic peptide (BNP) was measured with a bedside assay [AS Maisel et al. *Rapid Measurement of B-Type Natriuretic Peptide in the Emergency Diagnosis of Heart Failure*; *N Engl J Med* 2002; 347:161–167]. The clinical diagnosis of congestive heart failure (Gold Standard) was adjudicated by two independent cardiologists, who were blinded to the results of the BNP assay.

As it happens for all diagnostic tests, the accuracy of BNP to diagnose CHF varied based on what value of BNP was considered abnormal. For a BNP = 150 as the cutoff value, the Sensitivity was 85% and the Specificity was 83% (1-Specificity = 17%), as marked by the green arrow in Fig. 2.1.

That point in the curve of Fig. 2.1 is defined by the Sensitivity and Specificity generated by the number of False Negatives and False Positives below and above, respectively, of the cutoff value, which is depicted by the vertical line in Fig. 2.2

However, we may be unhappy with such low Sensitivity. In that case, we may decide to call abnormal any value greater than 100. That gives us a Sensitivity of 90% and a Specificity of 76% (1-Specificity = 24%) as marked by the red arrow in Fig. 2.3.

We got better Sensitivity because we reduced the number of False Negatives, and worse Specificity, because we increased the number of False Positives. That new cutoff value is depicted by the red vertical line in Fig. 2.4. If we keep lowering the value that we call Abnormal, the Sensitivity keeps improving at the expense

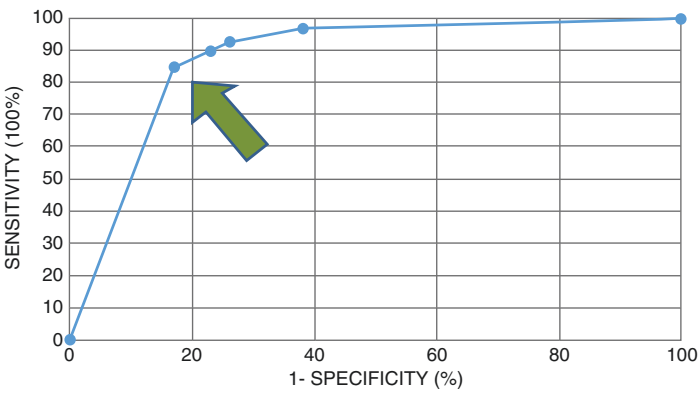


Fig. 2.1 ROC Curve. The arrow points at the Sensitivity and Specificity pair corresponding to a cutoff value of 150

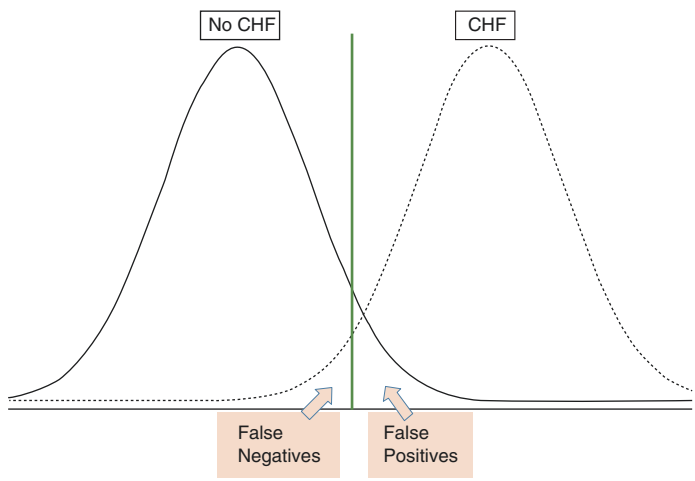


Fig. 2.2 BNP Distribution by Disease Status. The diagnostic cutoff of 150 is depicted by the vertical line. [CHF = Congestive Heart Failure]

of the Specificity. This happens because, as we have seen, there is an overlap in the distribution of values for people with and without the disease.

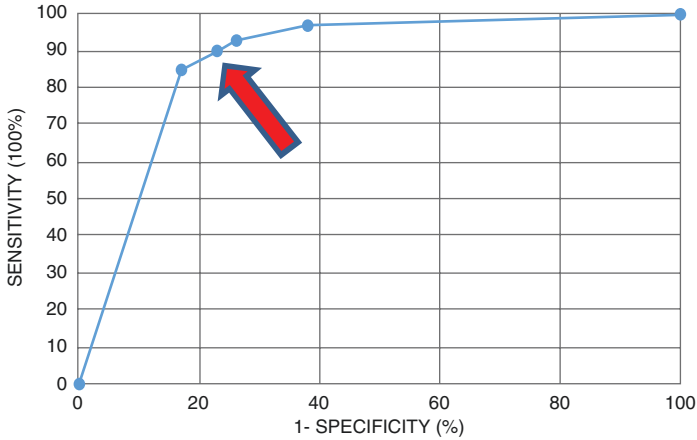


Fig. 2.3 ROC Curve. The arrow points at the Sensitivity and Specificity pair corresponding to a diagnostic cutoff value of 100. Sensitivity is now higher but Specificity is lower

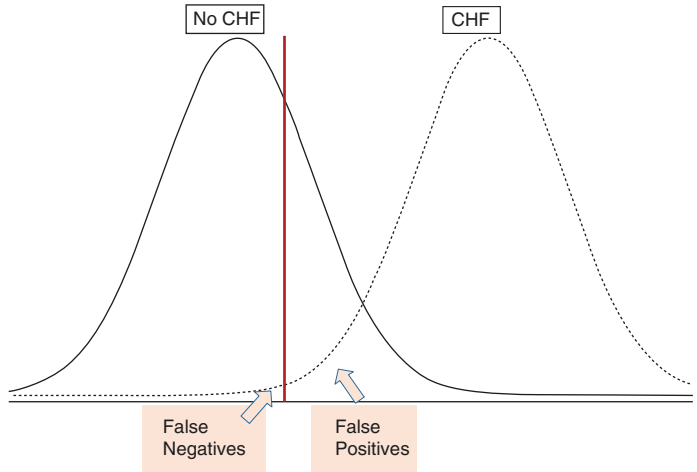


Fig. 2.4 BNP Result Distributions by Disease Status. The diagnostic cutoff value of 100 is depicted by the vertical line. As compared to a cutoff of 150, the number of false negatives is smaller but the number of false positives is larger. [CHF = Congestive Heart Failure]

When building an ROC curve we plot [1-Specificity] instead of Specificity in the vertical axis because that allows us to measure the Area Under the Curve (AUC). **The AUC reflects the ability of the test to discriminate between those who have the disease from those who do not have it. We call that the Discriminant Accuracy of a test. Of note, the AUC is equivalent to the C statistic for binary outcomes.**

An AUC of 0.50 means a test is useless—it is like tossing a coin to make the diagnosis. A test with an AUC > 0.70 is a good test, whereas an AUC of 0.80 or more is excellent, and an AUC > 0.90 corresponds to an outstanding test.

We can use the AUC to compare two diagnostic methods to each other in regards to their discriminant accuracy: the one with a significantly greater AUC has a better diagnostic performance. In Fig. 2.5 the test plotted in red has a larger AUC, and therefore

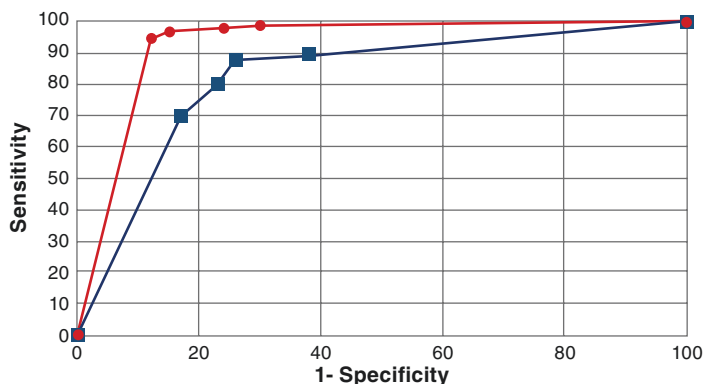


Fig. 2.5 Receiver Operating Characteristic Curves for Two Diagnostic Tests. The Test Depicted in Red Has a Significantly Larger Area under the Curve—It Has Better Discriminant Accuracy

it is a better test than the one plotted in blue. The P-Value tells us that the difference in AUC is statistically significant.

2.3.1 Effect of Threshold Changes on Predictive Values

Figure 2.6 depicts what happens to the Predictive Values when we lower the diagnostic threshold. As expected, we see the Negative Predictive Value improve with the Sensitivity, and the Positive Predictive Value decrease with the Specificity.

The opposite happens when we raise the value we consider abnormal: we increase the Specificity and the Positive Predictive Value, while we decrease the Sensitivity and the Negative Predictive Value (Fig. 2.7).

$$\uparrow \text{ Sensitivity} = \frac{\text{TP}}{\text{TP} + \downarrow \text{FN}}$$

$$\downarrow \text{ Specificity} = \frac{\text{TN}}{\uparrow \text{FP} + \text{TN}}$$

$$\downarrow \text{ Positive Predictive Value} = \frac{\text{TP}}{\text{TP} + \uparrow \text{FP}}$$

$$\uparrow \text{ Negative Predictive Value} = \frac{\text{TN}}{\text{TN} + \downarrow \text{FN}}$$

Fig. 2.6 Effect of Choosing a Lower Value as Diagnostic Threshold. Sensitivity and Negative Predictive Value increase, whereas Specificity and Positive Predictive Value decrease

$$\downarrow \text{ Sensitivity} = \frac{\text{TP}}{\text{TP} + \uparrow \text{FN}}$$

$$\uparrow \text{ Specificity} = \frac{\text{TN}}{\downarrow \text{FP} + \text{TN}}$$

$$\uparrow \text{ Positive Predictive Value} = \frac{\text{TP}}{\text{TP} + \downarrow \text{FP}}$$

$$\downarrow \text{ Negative Predictive Value} = \frac{\text{TN}}{\text{TN} + \uparrow \text{FN}}$$

Fig. 2.7 Effect of Choosing a Higher Value as Diagnostic Threshold. Sensitivity and Negative Predictive Value decrease, whereas Specificity and Positive Predictive Value increase

2.4 F-Score

The F-score (sometimes called the F-measure, or the F1 score) is an accuracy estimate frequently used in Artificial Intelligence and it is finding its way into the medical literature.

It is the harmonic mean of the Precision and Recall and it can be computed as:

$$\text{F-Score} = (2\text{PPV} * \text{Sensitivity}) / (\text{PPV} + \text{Sensitivity})$$

This may all sound rather fancy, but it has practical implications. The F-score increases or decreases as the prevalence rises or falls, respectively.

To see what that means let us examine an example. We have a test with 90% Sensitivity and 90% Specificity, but it is used in two samples with different prevalence for the outcome: Fig. 2.8 depicts how the PPV and the F-score fall when the prevalence drops, although the Sensitivity and Specificity remain the same.

SENSITIVITY	0.9	80% Prevalence					
SPECIFICITY	0.9						
		DISEASE POS		DISEASE NEG			
TEST POS		720	20	740		PPV	0.98
TEST NEG		80	180	260		F1	0.94
		800	200	1000			

		DISEASE POS		DISEASE NEG			
TEST POS		180	80	260		PPV	0.70
TEST NEG		20	720	740		F1	0.79
		200	800	1000			

Fig. 2.8 Effect of Prevalence on the F-score. The F-score for the same test drops from 0.94 to 0.79 because the Prevalence drops from 80% to 20%

The C-statistic (Area Under the Curve of the ROC curve), on the other hand, has the advantage of being relatively insensitive to prevalence changes.

2.5 Common Biases in the Evaluation of Diagnostic Tests

2.5.1 Spectrum Bias

A group of rheumatologists report to have identified a new serologic marker for Rheumatoid Arthritis (R.A.). They publish a study of that marker, which you are reviewing for publication in a journal. They measured their new marker in 300 patients with active, severe R.A. They also measured it in 300 age- and gender-matched controls, from their Diabetes clinic, who had no history of inflammatory arthritis. As a reviewer, what are your thoughts?

Well, you should worry that they sampled from the two extremes of the spectrum of R.A. disease severity: the cases with severe R.A. versus the controls who have absolutely no arthritic inflammation. That is likely to result in two well separated curves for the marker's levels, as seen in Fig. 2.9.

The lack of any overlap between the curves means that we will observe absolutely no False Positives or False Negatives, leading to a false impression of perfect sensitivity and specificity for the test.

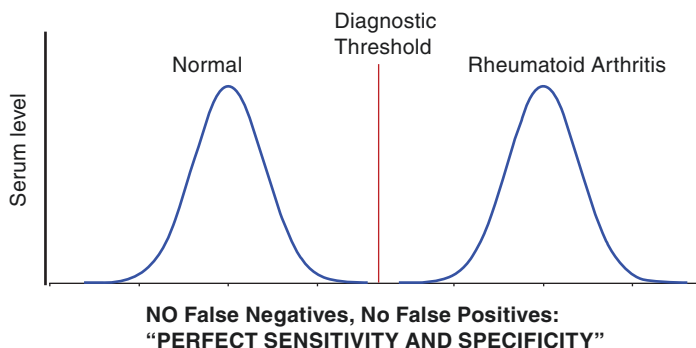


Fig. 2.9 Spectrum Bias. Test results for people with severe active disease naturally do not overlap with those for people without any rheumatologic complaints, causing a spurious finding of perfect sensitivity and specificity

This is an exaggerated example of Spectrum Bias. The participants in the study had a completely different spectrum of disease severity from what you will see in the mix of patients who come to your office with yet undiagnosed arthritis.

2.5.2 Post-Test Referral Bias

We see this type of bias when researchers perform a retrospective analysis of a clinical database. For example, a cardiology research network performs a retrospective study to assess the accuracy of Myocardial Perfusion Scintigraphy (MPS) for the diagnosis of Coronary Artery Disease (CAD). They use a clinical database from 20 centers that collected MPS and coronary angiogram data over the last 5 years. With the coronary angiogram as the “Gold Standard”, they report the following accuracy for MPS: Sensitivity = 95%; Specificity = 64%.

The researchers are baffled by the low specificity of MPS, as previous studies had reported specificities in the 90’s. What question should be asked at this point?

You must figure out what proportion of patients with a negative MPS actually had a coronary angiogram. If cardiologists really

trusted MPS it is quite possible that a tiny minority of patients with a normal (negative) MPS were referred for an angiogram. That situation is depicted in Fig. 2.10.

The result of this bias, known as Post-test Referral Bias in the Cardiology literature, is an underestimation of the test's specificity and an overestimation of its sensitivity, as depicted in Fig. 2.11. The small number of False Negatives in the denominator increases the estimated sensitivity, whereas the large number of False Positive in its denominator decreases the estimated specificity.

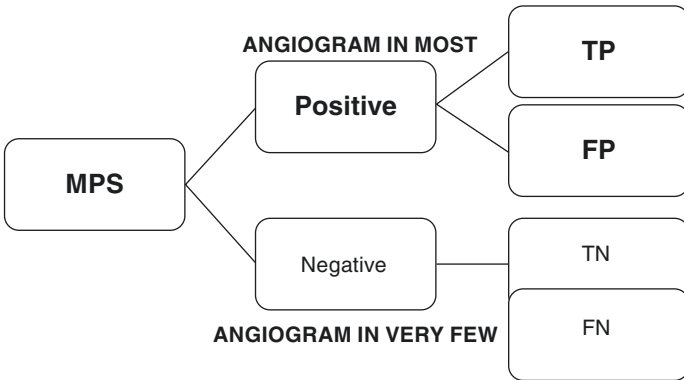


Fig. 2.10 Post-Test Referral Bias for a Trusted Test. Most patients referred for an angiogram had a positive Myocardial Perfusion Study (MPS). [TP = true positives; FP = false positives; TN = true negatives; FN = false negatives]

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$



$$\text{Specificity} = \frac{TN}{TN + FP}$$



Fig. 2.11 Post-Test Referral Bias for a Trusted Test. Sensitivity increases and Specificity decreases. [TP = true positives; FP = false positives; TN = true negatives; FN = false negatives]

Let us examine what could happen in the opposite situation, when clinicians did not trust a negative diagnostic result. Self et al. reported a cross sectional study of 3,423 adult patients presenting to 12 EDs in the United States from July 1, 2003 through November 30, 2006 who underwent both Chest X-Ray (CXR) and chest Computed Tomography (CT) for routine clinical care of suspected pneumonia (*Am J Emerg Med* 2103). They used the CT as Gold Standard to assess the CXR accuracy. The CXR test characteristics for detection of pulmonary opacities included: sensitivity 43.5% (95% CI: 36.4%–50.8%); specificity 93.0% (95% CI: 92.1%–93.9%).

Why was the sensitivity so low? The most likely explanation is that clinicians were suspicious of a normal CXR in patients presenting with all the symptoms of pneumonia but were satisfied with the CXR when they saw an infiltrate that confirmed their clinical suspicion.

That meant that most patients with negative CXR ended up having a chest CT, whereas a minority of patients with a positive CXR had the CT scan (Fig. 2.12). They had a lot of “negatives” in their sample but only a few “positives”. That falsely decreased the sensitivity, while it maximized the specificity, as shown in Fig. 2.13.

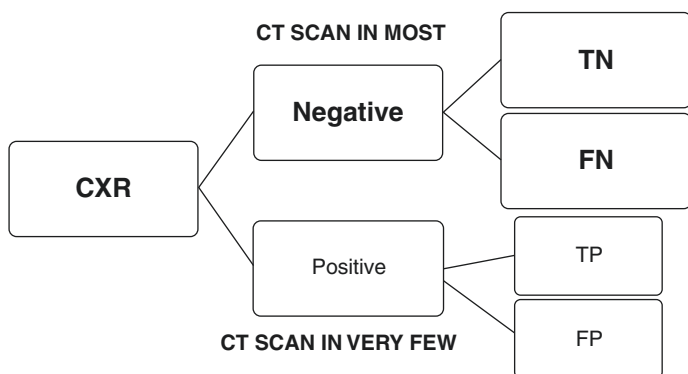


Fig. 2.12 Post-Test Referral Bias for an Untrusted Test. Most patients referred for a CT scan had a negative Chest X-Ray (CXR). [TP = true positives; FP = false positives; TN = true negatives; FN = false negatives]

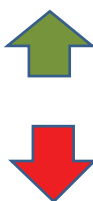
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$
$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$


Fig. 2.13 Post-Test Referral Bias for an Untrusted Test. Specificity increases and Sensitivity decreases. [TP= true positives; FP= false positives; TN= true negatives; FN=false negatives]

2.5.3 Biased Gold Standard

The most important feature of a Gold Standard is that it should not be contaminated by the diagnostic test we are trying to assess. The diagnostic test and its corresponding gold standard must be kept absolutely separated from each other.

Let us examine an example of how bias may be generated when the ascertainment of the Gold Standard is contaminated, somehow, by the test we are evaluating.

A network of European investigators wanted to know how accurate pro-BNP was for the diagnosis of CHF in their setting after using the test for about 4 years. They studied 500 consecutive patients attending clinics and/or the emergency department of 8 different hospitals. None of those patients had a prior diagnosis of CHF and they all presented with symptoms suspicious for CHF. They all underwent the same pro-BNP test, and the results were available to treating clinicians. The researchers had access to the complete medical records, so they took out any mention of pro-BNP values and gave the records to a panel of three independent cardiologists. Those cardiologists determined whether each patient had CHF or not. They ascertained the Gold Standard through consensus—they all had to agree on the final determination.

What was wrong with that study? It is quite possible that the pro-BNP results influenced the entire diagnostic and therapeutic course, because clinicians knew the pro-BNP value and applied it.

That would be reflected in the medical records, and the cardiologists ascertaining the Gold Standard could not avoid a clinical gestalt that dominated the narrative in the records. Researchers cannot avoid the influence of what clinicians thought and how they documented it, even after the BNP values are removed from the records. The test results are thus “confirmed” more often, making the test appear more accurate than it really is.

Of course, it would have been ethically impossible to deprive clinicians from the pro-BNP results, so it is essential to assess diagnostic tests before they become part of the standard of care. Once the test is already in clinical use a non-randomized outcomes research study can be very informative, as we discuss later.

2.5.4 Highly Selected Populations

We always prefer multi-site to single-site studies. When you recruit participants from a single site you may end up with a highly selected sample. The results from such a study may have very limited applicability elsewhere.

For example, Haydel and colleagues. [Haydel MJ et al. *N Engl J Med* 2000; 343:100–105] reported a very sophisticated study that developed and validated a set of clinical criteria to identify patients with minor head injury who do not need to undergo a head CT. First, they recorded clinical findings in 520 consecutive patients with minor head injury who had a normal score on the Glasgow Coma Scale and normal findings on a brief neurologic examination; the patients then underwent CT. Using recursive partitioning, they derived a set of criteria to identify all patients who had abnormalities on CT scanning. In the second phase, the sensitivity and specificity of the criteria for predicting a positive scan were evaluated in a group of 909 patients.

All patients with a positive CT scan had one or more of seven findings: headache, vomiting, an age over 60 years, drug or alcohol intoxication, deficits in short-term memory, physical evidence of trauma above the clavicles, and seizure. Among the 909 patients in the second phase, the sensitivity of the seven findings combined was 100 percent (95 percent confidence interval, 95 to 100 per-

cent). All patients with positive CT scans had at least one of the findings. Surprisingly, taking an oral anticoagulant was not one of the predictors of an abnormal head CT in their statistical model. Why not?

Their results reflected the unique sample recruited into the study. They performed a single-site study, recruiting participants only at a level-1 trauma center in New Orleans. Their patients were very young (mean age 36 years), and given the location of the Hospital, they suffered the head trauma at the French Quarter or its vicinity, while they were probably having too much fun. The study design was impeccable, but the results probably do not apply to the elderly patients with head trauma we evaluate at our Emergency Department in New York City. Many of our patients take oral anticoagulants, and that fact plays a major role in our decision to order a head CT scan.

2.6 Avoiding Biases

When you read a study addressing the development, validation, and/or improvement of prediction models for diagnosis or prognosis you should always start by making sure that the study complies with TRIPOD statement. [tripod-statement.org] This minimum set of rules provides a solid benchmark to ensure the validity of this type of study.

Figure 2.14 summarizes common biases and how to prevent them.

In summary, a study of a diagnostic/predictive test *should be prospective, with recruitment of a representative sample of patients who, as much as possible, should be exhibit the entire spectrum of disease severity. Recruitment should preferably take place at different locations, with diverse patient characteristics. Each participant should undergo both the diagnostic test and the gold standard determination independently, and in a completely blinded manner.*

UNBIASED SCENARIOS	BIASED SCENARIOS
Sample is representative of the population expected to have an indication for diagnostic testing.	Convenience sampling; single-site sampling.
Case mix from mild to severe disease forms of the disease spectrum, as expected in the clinical setting.	Only overt, severe disease versus very mild disease: Spectrum Bias.
Gold standard prospectively obtained in all participants.	Gold standard obtained by clinicians depending on diagnostic test result: Post-Test Referral Bias.
Test and gold standard are completely blinded to each other.	Contamination of gold standard by knowledge of test result.

Fig. 2.14 Unbiased and Biased Scenarios. The correct methodology listed on the left column prevents biases when assessing a diagnostic test

2.7 Checklist for the Assessment of a Diagnostic Test

- Who were the subjects being evaluated with the diagnostic test? Were they similar enough to your patients?
- Was there Spectrum Bias?
- Were the diagnostic test and the Gold Standard performed prospectively, independently from each other, and in all subjects?
- Was the diagnostic test being evaluated adequately described (e.g., manufacturer, device utilized)?
- Was the Gold Standard appropriate? Is there a better option?
- How did the diagnostic test perform? What were its sensitivity and specificity?
- Do you expect the Predictive Values estimated in the study to be similar to what you will have in your population? (Remember the influence of Prevalence on the Predictive Values).
- What can the Receiver Operating Characteristic (ROC) curve tell you? Is there an optimal cutoff point for a balance between sensitivity and specificity?

- How does this test compare to other diagnostic methods, if available? Did the study compare the area under the ROC curve for two similar methods?
 - Could we apply the results in our practice?
 - Are the test results and their interpretation reproducible in our setting?
 - Would the results be applicable in our setting; i.e. would the test performance be similar?
 - Would the test results change patient management?
-

2.8 Screening

We perform screening to identify patients with a condition before they have any symptoms. We do it because we expect that earlier diagnosis will lead to earlier treatment, which in turn will reduce morbidity and mortality.

An ideal set of circumstances for screening includes:

- **The condition should be relatively prevalent.** Please note the adverb “relatively”. Phenylketonuria is relatively uncommon but we screen for it within 48 hours of birth because we can avoid severe disease manifestations if we start early intervention.
- **The test is accurate.** At the very least, it should be highly sensitive, so all patients with a positive test result may undergo a subsequent more specific test.
- **If left untreated the condition causes substantial morbidity and/or mortality.**
- **An efficacious treatment is available, and earlier treatment results in better outcomes.**

If these conditions are not met, we may still perform screening when the diagnosis is valuable for counselling, risk prediction, and/or maximization of quality of life, or when it is requested by an informed patient who believes it would be highly contributory to their autonomous decision making.

You should beware of observational studies that claim a screening strategy is beneficial because it increases early detection of the condition and/or longer survival times after detection. Those two features without an accompanying reduction in age-adjusted case-specific mortality and all-cause mortality may just represent lead time bias and length bias, as we discuss in the Observational Studies chapter. Similarly, before and after studies of screening strategies may suffer from chronology bias.

Ideally, a new screening test or strategy should be assessed using a prospective design, with randomization to either undergo the new screening or not, and with a follow-up that is appropriately designed to capture effects on all outcomes of interest including both benefit (reduction in poor disease outcomes) and harm, including the psychological and physical effects of additional testing and/or treatment. A commonly used statistic that summarizes the benefit of screening is the Number Needed to Screen. This statistic has the advantage of capturing the entire sequence from early diagnosis to efficacious treatment that results in the avoidance of poor health outcomes.

Number Needed to Screen: the number of people who need to undergo screening to prevent the occurrence of one bad outcome.

Use of a Diagnostic Test

3

3.1 The Two-by-Two Table in Different Scenarios

Let us examine an example to see how the two-by-two table allows us to implement the Bayesian approach to diagnosis. You are back to an ED rotation after an extremely short vacation. Your first patient is a very sweet 73-year-old lady who was brought by her family because of sudden onset shortness of breath. She suffered a fall with a right hip fracture 2 weeks ago, for which she underwent surgical repair. Because of a recent severe gastrointestinal hemorrhage, she has not received any anticoagulants after the surgery. Her only other medical history is well controlled hypertension. She is up to date with her cancer screening. When you examine her she denies any pain since her hospital discharge. She is tachycardic, with a heart rate of 105 beats per minute. Her exam reveals complete normal lungs and heart, and no evidence of deep venous thrombosis of her lower extremities.

You believe pulmonary embolism (PE) may be the reason for the dyspnea, and decide to look up the Geneva Criteria, and determine that her score is eight points (Table 3.1).

The 8 points are consistent with approximately 30% prevalence of PE in patients like her. In other words, she has a 30% Pre-Test Probability of PE. How do we enter that information in a

Table 3.1 Score from Geneva Criteria

Criterion	Points
Age > 65	1
Previous DVT or PE	3
Surgery or fracture within 1 month	2
Active malignancy	2
Unilateral limb pain	3
Hemoptysis	2
HR 75–94	3
HR > 94	5
Pain on lower limb deep vein at palpation AND unilateral edema	4

Table 3.2 Two-by-Two Table with 30% Prevalence

	Disease present	Disease absent	
Abnormal test	TP	FP	
Normal test	FN	TN	
	300	700	1000

2-by-2 table? We start with an arbitrary size of 1000 for the table, as can be seen in Table 3.2.

We order a D-dimer test for her, which at a pre-specified threshold has a reported Sensitivity of 95% and Specificity of 45%. How do we use the 2-by-2 table to estimate her Post-test Probability of having PE?

We remember that $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$, so we know that the TP are $(300 * 0.95) = 285$.

The number of FN is calculated as: $300 - 285 = 15$.

Similarly, we know that $\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$, so we know that the TN are $(700 * 0.45) = 315$.

The number of FP is calculated as: $700 - 315 = 385$.

We proceed to fill in the cells (Table 3.3):

Let us see what happens if our patient had an ABNORMAL test. In that case she would belong to the first row from the top, and we would like to estimate the probability that hers is a True Positive result.

Table 3.3 Two-by-Two Table for 30% Pre-Test Probability, 95% Sensitivity, and 45% Specificity

	Disease Present	Disease Absent	
Abnormal test	TP	FP	
	285	385	670
Normal test	FN	TN	
	15	315	330
	300	700	1000

In other words, we want to calculate the PPV = $TP / (TP + FP) = 285/670 = 0.42$.

A 42% Post-test Probability is probably high enough to offer anticoagulation, at least while we explain the findings and ask the patient about her preferences. Most patients would prefer to have a higher post-test probability before agreeing to a treatment like anticoagulation. In that case a second test, like a CT angiogram, would be warranted.

Let us review the opposite scenario. If she had a NORMAL test result, what would have been her Probability of having PE?

In that case, she would have belonged to the bottom row in the table, and we would like to know the probability of that normal result being a False Negative. That probability can be estimated as:

$$FN / (FN + FP) = 15 / 330 = 0.045.$$

The Negative Predictive Value would be 95.5%. That would mean that there is a 95.5% probability that our patient does not have PE. We should consider other diagnoses to explain her dyspnea.

A test with high sensitivity but lower specificity should be optimally ordered for patients with low or low-intermediate pretest probability, because a negative test would allow us to “rule out” the disease, and move on with our diagnostic evaluation.

3.2 Pretest Probability Estimates

An optimal approach to diagnosis requires an active effort to find in the literature the least biased estimates of the test's sensitivity and specificity, as well as validated methods to estimate the pretest probability in populations similar enough to our patient. Many well designed studies have provided us with means to estimate the pretest probability of conditions such as pulmonary embolism, and coronary artery disease, based on clinical features upon presentation.

It is essential to always consider the Pre-test Probability before we make the decision to order a diagnostic test. In particular, when the pretest probability is either too low or too high the value of further diagnostics testing is very low.

If a patient has an extremely low Pre-test Probability of having a disease, no test for that disease should be ordered, because an alternative diagnosis should be considered.

If a patient has an extremely high Pre-test Probability, treatment should be initiated without further delay. Waiting for another test would be unwise.

3.3 Likelihood Ratios and Fagan Nomogram

The two-by-two table, as we have just discussed, provides a user-friendly and intuitive approach to estimate Disease Probability when we know the Pretest Probability for our patient and the Sensitivity and Specificity of the test. This is nothing else than a tabulated application of Bayes Theorem.

There is another simplified approach to apply Bayes Theorem to diagnosis. It consists of the use of a nomogram that takes as inputs the Likelihood Ratios and the Pretest Probability. This

nomogram is known as Fagan nomogram, and it was first published in the 1970s in the New England Journal of Medicine. The left column are pretest probabilities, the middle column are likelihood ratios, and the right column are posttest probabilities (Fig. 3.1). All we have to do is draw a straight line through our

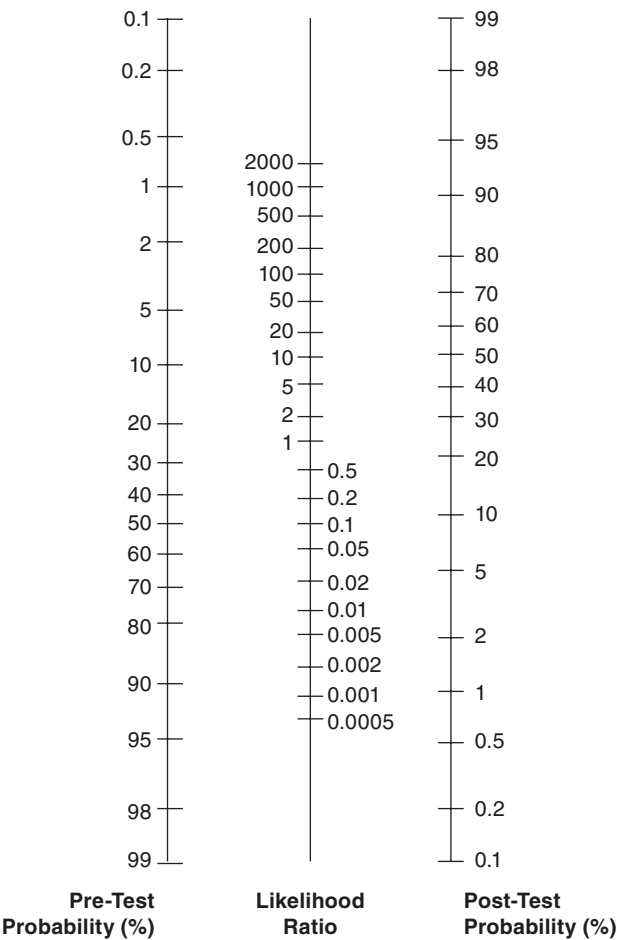


Fig. 3.1 Fagan Nomogram. Drawing a straight line through the pre-test probability and the likelihood ratio gives you the post-test probability

pretest probability and likelihood ratio to obtain the estimated posttest probability.

The nomogram carries out the following calculations for you, according to Bayes Theorem:

1. It calculates Pretest Odds = Pretests Probability / 1 – Pretest Probability.
2. It multiplies Pretest Odds * Likelihood Ratio, which gives the Posttest Odds.
3. It calculates Posttest Probability = Posttest Odds / 1 + Posttest Odds.

Of note, there are innumerable online calculators and apps that work just like Fagan Nomogram, estimating posttest probability through Bayesian synthesis of the pretest probability and likelihood ratios.

3.4 Use of Predictive Models

Outcome prediction is in many ways very similar to the diagnostic process. An accurate prognosis is to the future what an accurate diagnosis, with risk stratification, is to the present. During the COVID-19 pandemic it was extremely important to decide when it was safe to discharge a patient home after they improved. Razavian and colleagues conducted a very elegant study to develop and validate a parsimonious regression model that estimates the probability a patient will have a favorable outcome if discharged. *A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients.* N Razavian et al. *npj Digital Medicine* (2020)3:130; <https://doi.org/10.1038/s41746-020-00343-x>

Razavian et al. defined favorable outcome as the absence of all of the following poor outcomes within 96 hours: (1) Death or discharge to hospice; (2) ICU admission; (3) Significant oxygen support (mechanical ventilation, non-invasive positive-pressure ventilation including BIPAP and CPAP, high-flow nasal cannula,

face mask or nasal cannula flow rate greater than 6 L/min); (4) If discharged, re-presentation to the emergency department or readmission.

The resulting predictive model was efficient at diagnosing who would do well after discharge. They were able to integrate the model into their clinical workflow, and shared their coefficients and intercept, so we can all apply their findings using simple tools, like Excel spreadsheets. The next figure shows how such a spreadsheet works. In this case, it requires entering 0 or 1 for dichotomous variables, and the actual value (in the units used by the authors) for continuous variables. The example below shows the model predicting that a patient with thrombocytopenia and elevated C-reactive protein is quite unlikely to do well if discharged at that point (they would have only a 22% probability of a favorable outcome). It is probably best to postpone discharge until the platelet count improves and the C-reactive protein levels are lower (Fig. 3.2).

Like all predictive models, the one we have just discussed should be reassessed and recalibrated as needed, in different populations. In addition, a careful collective decision is needed to determine what estimated probability of favorable outcome is high enough to use as a site-specific threshold for discharge, at any given time.

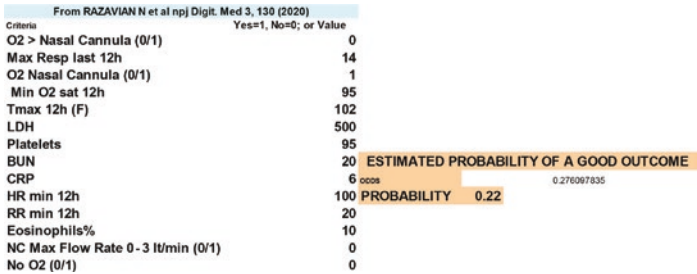


Fig. 3.2 Prediction of Good Outcome After Discharge. This patient has only a 22% probability of doing well if discharged now



Observational Studies

4

4.1 Observational Studies. General Considerations

In Medicine we can divide studies of the association between Exposures and Outcomes into two large categories:

- Interventional: we intervene, exposing patients to different treatments in a Randomized Clinical Trial. This gives the strongest evidence we can obtain about causation.
- Observational: we identify a sample of our population and observe their exposures and outcomes.

Observational studies are the only possibility when the exposure either cannot be administered (educational status) or would be unethical to administer (tobacco use). They are also the only practical approach when the outcome is either very rare or takes a long time to occur. We can classify Observational Studies into the following groups:

- Cross-Sectional
- Case-Control
- Longitudinal Cohort.

When performing or reading observational studies we need to consider the possibility of Confounding and Bias.

Confounders are variables associated with the exposure, that are also associated with the outcome without being a mediator, and should be measured and controlled for in the analysis.

Unmeasured or poorly measured confounders can severely compromise the validity of our observational findings.

Bias is a fundamental flaw in the design of the study that affects the results, and cannot be controlled, or adjusted for.

Before we review each type of observational study, it should be noted that all of them share the same important limitations, as follows.

A major limitation is that the exposure was not randomly allocated, and there is always a possibility that the association we found may be explained by confounders. You may argue that if potential confounders are well measured they can be adjusted for in the analysis. However, there are frequently unmeasured, poorly measured, or even unknown confounders. They generate residual confounding after all our adjusting.

Observational studies only assess associations and do not provide definite proof of a causal relationship. For example, two variables may be associated to each other because they share a preceding cause that we have not identified yet.

**There can always be Residual Confounding.
Association does not prove Causation.**

In addition to those universal caveats, specific biases may arise depending on the type of study. Regardless of the type of observational study you read, you should begin your assessment of their quality and validity by making sure that they comply with the STROBE reporting guidelines. [strobe-statement.org] The STROBE checklist is very useful to both reviewers and readers. You should keep in mind that publication in a prestigious journal does not guarantee compliance with basic validity standards, as shown by the work of Aghazadeh-Attari and colleagues. [Aghazadeh-Attari, J., Mobaraki, K., Ahmadzadeh, J. et al. *Quality of observational studies in prestigious journals of occupational medicine and health based on Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: a cross-sectional study. BMC Res Notes* **11**, 266 (2018). <https://doi.org/10.1186/s13104-018-3367-9>].

4.2 Cross-Sectional Studies

This is the simplest approach. It takes a snapshot picture of a sample, and describes the exposures and outcomes at that point.

There was no attempt to collect all data about exposures before the outcome took place.

Because of their limited scope cross-sectional studies are the weakest type of observational study. They do not assess the temporal relationship between cause and effect, and thus they may suffer from *reverse causality bias*. For example, the old cross-sectional studies showing an association between lower socioeconomic status and schizophrenia were highly suspect of reverse causality. Did poverty increase the risk of schizophrenia, or did schizophrenia increase the risk of becoming poor?

In addition, negative selective pressure from the exposure may bias the perceived association with the outcome in cross-sectional studies. For example, consider patients with uncontrolled diabetes. They have a higher risk to develop end stage renal disease (ESRD), but they also have higher cardiovascular mortality than those without diabetes. Their higher cardiovascular mortality will reduce the observable proportion of patients with uncontrolled

diabetes in a cross-sectional sample of ESRD patients. A cross-sectional study may therefore underestimate the societal importance of diabetes mellitus as a cause of ESRD.

Furthermore, if we study two variables that are associated because they share a common cause, but that shared cause is unknown to us, we may erroneously interpret the association between those variables as one them causing the other.

4.3 Odds Ratio

Because there is no prospective ascertainment of events in a cross-sectional design, the risk of incident events over time is not determined, and it is not appropriate to use the relative risk. Instead, we evaluate the association between exposure and outcome using the odds ratio. As discussed in the Survival Statistics Module of this notebook, we can calculate odds ratios for a single exposure in a univariate approach using the two-by-two table, or for multiple exposures using the multiple logistic regression model. In the case of a single exposure the univariate approach is to simply perform the cross-product ratio, using the product of “true cells” as numerator and the “false cells” in the denominator. We define “true” as being compatible with the association between the exposure and the outcome.

4.4 Case Control Studies

When a disease is rare and/or it takes a long time to go from the exposure to the outcome, we may not be able to perform a longitudinal cohort study. In that situation, we frequently perform a case-control study.

We begin identifying a group of people who have the disease or outcome: they are our cases. After that, we find a group of people who are as similar as possible to the cases in every way, but do not have the disease or outcome. They are our controls. Then, we look back in time in both groups to assess all exposures.

That is depicted in Fig. 4.1.

Start with the Disease status.

Look back to a time BEFORE the Disease for Exposures.

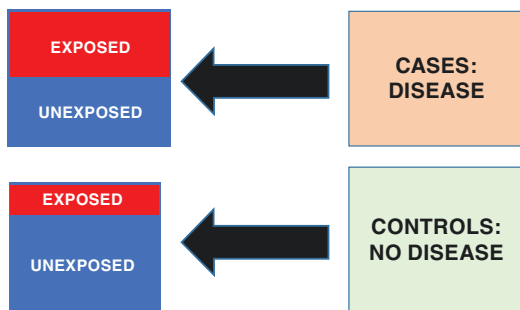


Fig. 4.1 Case-Control Studies. First, we identify people with the condition (cases) and without it (controls); we then proceed to determine who was exposed

A case-control design is efficient because it requires a smaller sample size than a cohort study, and it is more cost-effective.

Just like in cross-sectional studies it would be inappropriate to use the relative risk statistic in case-control studies. Instead, we use the odds ratio to measure the association between exposure and outcome.

4.4.1 Importance of the Control Group

The process of identifying and characterizing our control group is essential. In general, we want the controls to be as similar as possible to the cases, except for the presence of the outcome. We try to ensure that through Matching.

Matching is the process to select controls based on characteristics that make them very similar to our cases, except for the outcome. It prevents confounding by those characteristics. Obviously, you also lose the ability to assess any causal relationship between matched characteristics and the outcome.

Matching can be performed on a one-to-one basis, creating matched pairs of cases and controls. But we can also match the controls to the cases as groups, through frequency matching or stratification. For example, frequency matching by age ensures that cases and controls have a very similar mean age.

At the end of the day, we want our controls to represent the totality of our patients. It is always a good idea to examine the control group you created through matching to make sure they are not too healthy or too sick, as compared to your average patient.

Next, We will discuss biases commonly encountered in Cross-Sectional studies.

4.4.2 Ascertainment or Diagnostic Bias

Ascertainment or Diagnostic Bias takes place when the exposure made it more likely that we diagnosed (ascertained) the outcome.

For example, let us say we use radiology records to identify cases of ankylosing spondylitis, and we match them to controls from the same zip codes, who are of the same age and gender. After that, we go back to look for a history of neck trauma in all of them. What could go wrong?

Well, it is quite possible that many people after suffering neck trauma had an x-ray that gave us the incidental finding of ankylosing spondylitis. In other words, our exposure (trauma) increased the probability of ascertaining the outcome (ankylosing spondylitis).

4.4.3 Recall Bias

Case-control studies are particularly vulnerable to recall bias because we ask subjects about possible exposures after the outcome already took place—the longer the time after the outcome, the greater the risk of recall bias. This is understandable because people who have had an outcome (the cases) are much more likely to have thought about potential exposures. They probably ruminated about what caused the outcome, asking themselves “What

went wrong?” In turn, people without the outcome (the controls) had no reason to think about potential exposures.

We try to avoid recall bias by obtaining, as much as possible, our information about exposures from sources collected before the outcome took place. For example, we can look for exposure to pharmaceuticals from a prescription database that captured medication usage before the outcome.

If a longitudinal cohort study has captured information about the exposures and outcomes you are interested in, you should perform a nested study because it will avoid recall and ascertainment biases.

4.4.4 Interviewer Bias

Interviewer bias may take place when the researcher performing the interview is aware that a participant had the outcome. That knowledge may prompt the researcher to look for the exposure more exhaustively, asking more questions repeatedly and deviating from the interview script to probe deeper for the exposure.

4.4.5 Nested Case-Control Studies

We like the idea of using an existing database from a Cohort study to identify Cases and Controls within that cohort, and then go back in time to identify the exposures. This is a very time-efficient method, and we call it a Nested Case-Control Study, because it is nested within the cohort.

If the cohort has been well characterized in a prospective manner, we will have access to data of excellent quality. A nested approach addresses many limitations of non-nested analyses.

For example, old (non-nested) case-control studies reported an association between homocysteine levels and the risk of coronary artery disease. However, no association was found when nested case-control analyses were performed using high quality data from prospective studies (American Heart Journal; vol. 146, Issue 4, October 2003, 581–590).

Nested case-control studies have the following advantages:

1. We avoid Recall Bias because data about exposures were collected before the outcome happened.
2. The independent assessment of exposure and outcome prevents Ascertainment (or Diagnostic) Bias.
3. The approach is more cost-effective because the cohort data already exist.

4.5 Prospective Cohort Studies

Prospective cohort studies are the gold standard in observational epidemiology. They take a careful approach to sampling from the population, characterize their cohorts as best they can, and follow them up, hopefully for many years, to ascertain incident cases of the outcome. Naturally, we start with a group of people who have not had the outcome yet and we carefully assess them for all our exposures of interest. After that we follow them up to determine who gets the disease (Fig. 4.2).

One of the advantages of prospective cohort studies is that we can be ambitious when characterizing our cohort, and capture detailed information about many possible exposures. We can also save specimens such as blood samples, for future use. In addition, we can watch our cohort for many different outcomes.

Because the cohort does not have the disease at the beginning we are able to ascertain the incidence of the disease.

The association between exposure and outcome can then be evaluated using statistical methods that compare outcome risk between exposed and unexposed, such as the Relative Risk and the Hazard Ratio. In both cases, the change in absolute risk can be measured with the Number Needed to Harm.

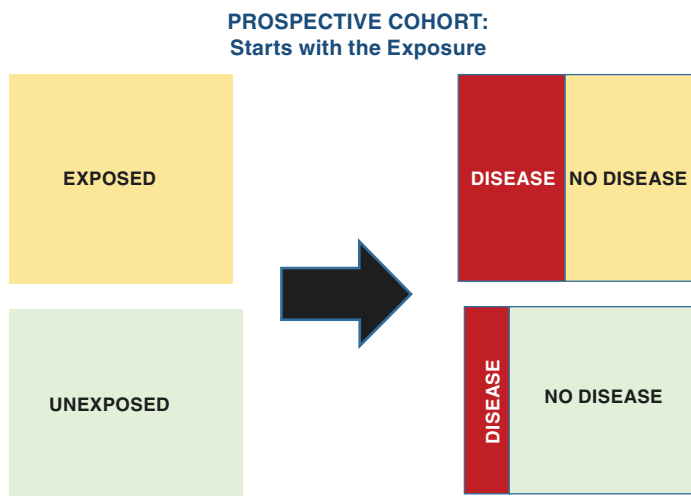


Fig. 4.2 Prospective Cohort Studies. First, we determine who has been exposed in our cohort; we then follow them up to determine who gets the disease

There are many possible biases in cohort studies, and we will discuss the most common of them next.

4.5.1 Selection Bias

Selection bias can take place at all stages of cohort studies, from recruitment to follow up. Selection bias happens during recruitment if the sampling method does not generate a sample that is representative of our population. For example, recruitment from a highly specialized medical center that sees many referrals will give you a sample that is sicker than the general population.

A major type of selection bias during follow up happens when the attrition is associated with the outcome. Those lost to follow up are taken out of the analytic risk pool from that point on—we censor them. If the risk of being censored is associated with the outcome, we have informative censoring. Informative censoring is not unique to observational studies, and it can be a problem in

randomized clinical trials as well. For example, in weight loss studies those participants who have gained back a lot of weight may be reluctant to return for a follow up visit.

When informative censoring affects quality of life studies that constitutes a type of Respondent Bias. For example, patients who have a poor quality of life and are depressed may be less willing to answer questions at follow-up than those who are happy.

A very similar type of selection bias, which can start at the recruitment phase of the study, is the Healthy Volunteer Bias. Health-conscious people volunteer more frequently to participate in certain studies, and do better at attending follow up visits.

4.5.2 Confounding by Indication (Prescription Bias)

In an observational (non-randomized) comparison of patients who received a given treatment to patients who did not, it is crucial to remember that those two groups often do not have the same underlying characteristics, and that may be the precise reason why they got different treatments.

Confounding by indication is the bias that arises in the observational study of treatment effects, and stems from the differing baseline risks, co-morbidities, and prognostic factors between patients who receive that treatment and those who do not receive it.

If a greater risk of the outcome causes a more frequent prescription of a treatment, the occurrence of the outcome may be wrongly attributed to that treatment in unadjusted analysis.

We frequently deal with confounding by indication using Propensity Scores.

For example, our colleagues Joshua Geleris et al. examined the association between hydroxychloroquine use and the risk of intubation or death in COVID 19 patients hospitalized at Columbia University Medical Center in 2020.

Treating clinicians had prescribed hydroxychloroquine in non-randomized fashion. They were likely to preferentially treat those patients who were sicker with hydroxychloroquine. That generated confounding by indication as shown in Fig. 4.3.

Geleris et al. adjusted their analysis using a Propensity Score. Propensity Scores are usually generated using a multiple logistic regression model. In this case the logistic regression model estimated the probability that any given patient received hydroxychloroquine based on their baseline characteristics. That estimated probability was the Propensity Score.

Logistic regression models are used because dichotomous outcomes cannot be analyzed with linear regression. Instead, we examine the association of any predictor/s with the log-transformed odds (logits) of the outcome.

The regression plot in Fig. 4.4 shows how we examine the value of Predictors in regards to their association with the Logit

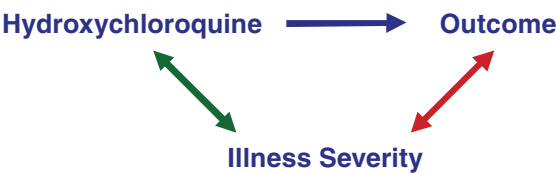


Fig. 4.3 Confounding by Indication. Illness severity is independently associated with the exposure and with the outcome

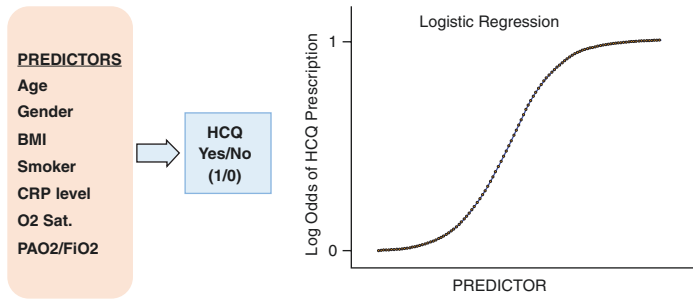


Fig. 4.4 Multiple Logistic Regression with Hydroxychloroquine Use as Outcome. The predictors are listed on the left hand side

for the outcome. Let us break that down: Odds are $[\text{Probability}/1 - \text{Probability}]$ and the log of the odds is known as a Logit, which explains the name of “Logistic” we give to this type of regression. The main output is an Odds Ratio for each exposure. I am sure you have always wanted to know all this, so you could bring it up during the first date with your future spouse.

If we did a good job at identifying the baseline characteristics that prompted clinicians to prescribe hydroxychloroquine in sicker patients, we can adjust for that propensity and thus remove (at least in part) the confounding caused by sicker patients (who naturally fared worse) receiving hydroxychloroquine more often. In other words, we use the Propensity Score to deal with confounding by indication.

However, how do we know that the Propensity Score did work appropriately? The ability of the Propensity Score to discriminate between those who actually received or did not receive the medication can be assessed using the C statistic, which is nothing else than the area under a Receiver Operating Characteristic Curve (see Diagnosis Module). It measures the accuracy of the propensity score to discriminate between those who ended up receiving hydroxychloroquine from those who did not. A C statistic of 0.5 suggests complete absence of diagnostic value, whereas a discriminant accuracy greater than 0.7 suggests good discriminant accuracy, and if it is greater than 0.8 it suggests excellent discriminant accuracy. In this study of hydroxychloroquine the C statistic for the Propensity Score was 0.81.

What impact did the Propensity Score have in the analyses ran by Geleris and colleagues?

The crude unadjusted analysis showed that patients who had received hydroxychloroquine were more than twice more likely to have a primary end-point event (hazard ratio, 2.37; 95% CI, 1.84 to 3.02). However, after adjustment with the propensity score there was no significant association between hydroxychloroquine use and intubation or death (hazard ratio, 1.04, 95% confidence interval, 0.82 to 1.32). The Propensity Score adjustment got rid of the confounding by indication. Without the adjustment we may have believed that hydroxychloroquine had a severe negative clinical effect. With the adjustment we see that there was no significant effect, neither harm nor benefit, from hydroxychloroquine.

Propensity Scores can be used at each stage of an observational study: sampling (matching those who received the drug to those who did not by their estimated propensity), or analysis (stratifying by categories of the score or using it as a covariate for adjustment). Geleris et al. used it in both ways, with very similar results.

In one of their adjusted analysis, they performed something called Inverse Probability Weighting. It sounds intimidating, but the concept is actually rather simple. Let us say we have two patients who received hydroxychloroquine, Mr. Smith and Mrs. Jones. Sadly, both died. Mr. Smith was elderly, obese, had a high C-reactive protein level on admission, and 86% O₂ saturation on room air. His risk was high and not surprisingly his Propensity Score was estimated as 0.92. Mrs. Jones was quite younger and thinner, her C-reactive protein level was 8 and her O₂ saturation on room air was 89%. Her Propensity Score was estimated as 0.36. When we use the Propensity Score for Inverse Probability Weighting to assess the association between hydroxychloroquine and mortality, the event observed for Mrs. Jones is given much greater value (is weighted much more) than the event observed for Mr. Smith. In other words, the weight given in the analysis to an event happening to a patient is inversely proportional to the Propensity Score for that patient (Fig. 4.5).

The ultimate validation of an observational study is provided when a Randomized Clinical Trial (RCT) reaches a similar conclusion. That was, indeed, the case for hydroxychloroquine as treatment for COVID-19 infection. The RECOVERY trial confirmed that there is neither benefit nor clear harm from hydroxychloroquine in this setting [*medRxiv* 2020.07.15.20151852; <https://doi.org/10.1101/2020.07.15.20151852>]. In RECOVERY 418 (26.8%) patients allocated to hydroxychloroquine and 788 (25.0%) patients allocated to usual care died within 28 days (rate

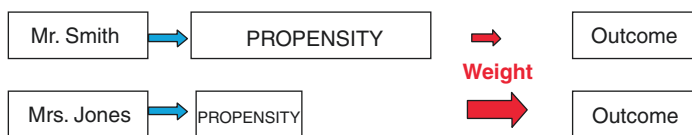


Fig. 4.5 Inverse Probability Weighting. The weight we give an outcome is inversely related to the propensity score for that person

ratio 1.09; 95% confidence interval [CI] 0.96 to 1.23; $P = 0.18$). This is compatible with the findings of the Geleris et al. observational study.

4.5.3 Immortal Time Bias

Immortal Time Bias occurs when investigators erroneously treat subjects as “ever exposed” when they analyze the time to events, assuming the exposure was present at the beginning of the follow-up, but it was not. The time prior to the exposure obviously preceded the outcome, and was thus “immortal” in regards to the outcome. The outcome could not have taken place during that time.

Let us consider the example with the follow up data from just two patients, as shown in the figure. We started following them up at the same time, and they both died during follow up. One of them was exposed to a potentially toxic drug—that exposed time, or time under treatment, is shown in red in Fig. 4.6.

We wonder whether the drug caused the death of the person exposed to it. If we (wrongly) analyze the person who took the drug as “ever exposed”, we (wrongly) change her entire follow up time into “Time under treatment” as seen in Fig. 4.7—their entire follow up time is depicted in red, as though the exposure had happened prior to the start of follow up. Doing this makes it wrongly seem as though the drug actually increased the survival time.

To avoid this bias, all time prior to the exposure should be counted as belonging to an unexposed subject, who never had the event, and whose follow up was interrupted (censored) at the time of the exposure. This treats the person who received the drug as two different people, before and after the exposure, as seen in Fig. 4.8.

This method makes it clear that people die faster after being exposed to the drug (the red bar is clearly shorter than the two yellow ones). In other words, this approach prevents immortal time bias in the analysis.

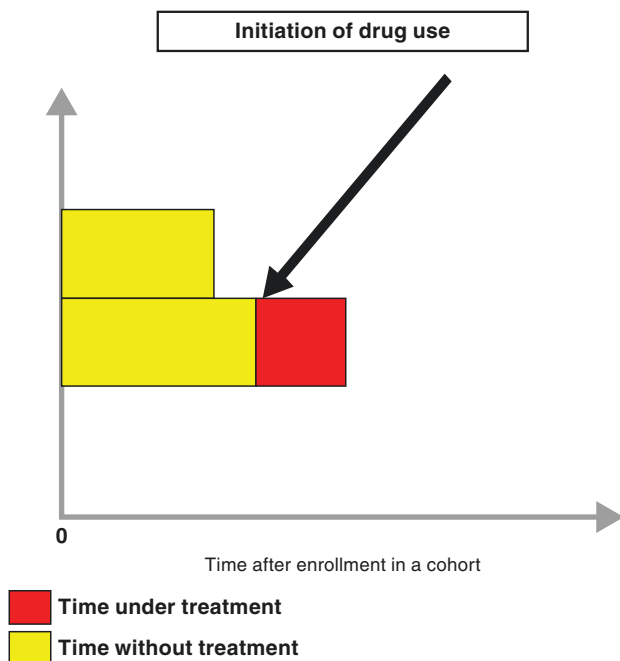


Fig. 4.6 Immortal Time Bias. Bars depict survival times; only one person was exposed to the drug. The time under exposure is in red

4.5.4 Attrition of those Susceptible

When a treatment causes harm shortly after being started, we may not detect that harm if we perform our study after those susceptible to the effect already had the outcome.

For that reason, we can get biased results when we compare a new drug to another that has been in the market for much longer, even if both have the same adverse effect.

Let us see an example. Patients with arthritis often feel better when they take COX-2 inhibitors, but the risk of cardiovascular (CV) events is a major concern with those drugs. A study in a

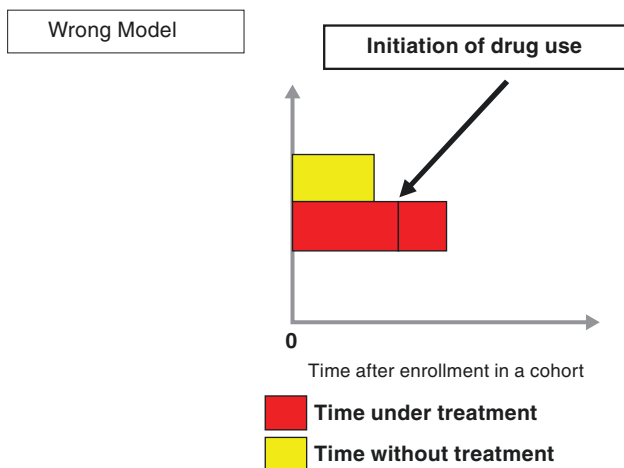


Fig. 4.7 Immortal Time Bias. The entire follow-up time for the person exposed is erroneously analyzed as under the influence of the exposure

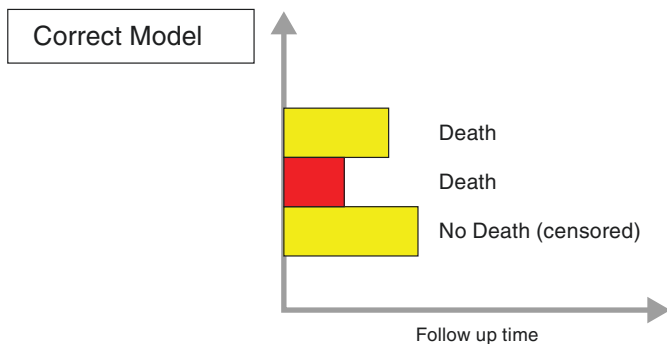


Fig. 4.8 Avoiding Immortal Time Bias. The time under the influence of the exposure is analyzed as corresponding to third person. That shows the negative effect on survival very clearly

national health system examined the risk of incident CV events in three major cities from 2010 to 2015, in people 65 or older who took one of two COX-2 inhibitors available through their coverage.

Taking the one available since 2010 was reportedly associated with higher CV risk than taking the one available since 2008. Of

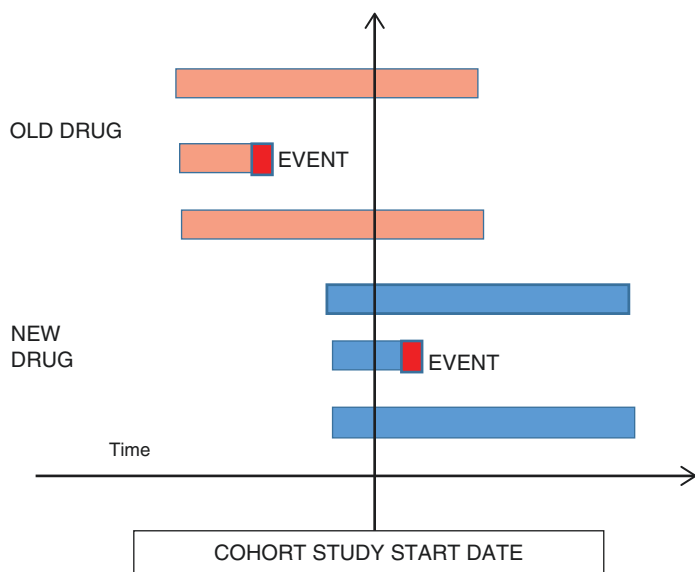


Fig. 4.9 Attrition of Those Susceptible. Event rates are identical in both groups but the study's timing allows event detection only in persons receiving the new drug

note, most CV events took place within 90 days of starting the drug. Fig. 4.9 shows a likely explanation for the observed difference between the drugs. Both drugs caused the event in 1/3 of patients early after starting it, but we observed only the event in the newer drug group (blue bars). The event in the older drug group took place before we started our study. Attrition of the susceptible patient took place in the older drug group before our study began.

4.5.5 Protopathic Bias

Protopathic bias occurs when a treatment is prescribed for an early manifestation of a disease that has not yet been diagnosed. When the disease is later diagnosed, a causal relationship maybe

wrongly inferred between the treatment and the disease. For example, headaches occur for some time before an imaging study is done and a brain tumor is diagnosed. People take medications for the headache. One may thus see an association between greater acetaminophen use and a subsequent diagnosis of brain tumors.

4.5.6 Chronology (Secular) Bias

Some studies have used historical controls to assess whether the introduction of a treatment had a positive effect on outcomes. That type of comparison between chronologically different cohorts should be avoided. From diagnosis to therapy there are secular trends leading towards better outcomes, and attributing that phenomenon to a single therapeutic exposure may simply constitute wishful thinking.

4.5.7 Non-Randomized Outcomes Studies

When we have already successfully adopted into our clinical practice a sequence of diagnostic and therapeutic methods into an algorithm to manage a serious illness it will be virtually impossible to perform a randomized trial to assess those methods. There is, however, a type of study that could confirm whether that algorithm, or variant thereof, is indeed safe and efficacious. Let us examine the Christopher study, which was a great example of how to conduct a non-randomized study in those circumstances [*van Belle A and the Christopher Study Investigators. Effectiveness of managing suspected pulmonary embolism using an algorithm combining clinical probability, D-dimer testing, and computed tomography. JAMA. 2006 Jan 11;295(2):172–9. <https://doi.org/10.1001/jama.295.2.172>. PMID: 16403929*].

The Christopher study was a prospective cohort study of consecutive patients with clinically suspected acute pulmonary embolism, conducted in 12 centers in the Netherlands from November 2002 through December 2004. The study population of 3306 patients included 82% outpatients and 57% women. Patients

were categorized as “pulmonary embolism unlikely” or “pulmonary embolism likely” using a dichotomized version of the Wells clinical decision rule. Patients classified as unlikely had D-dimer testing, and pulmonary embolism was considered excluded if the D-dimer test result was normal. All other patients underwent CT, and pulmonary embolism was considered present or excluded based on the results. Anticoagulants were withheld from patients classified as excluded, and all patients were followed up for 3 months. Pulmonary embolism was classified as unlikely in 2206 patients (66.7%). The combination of pulmonary embolism unlikely and a normal D-dimer test result occurred in 1057 patients (32.0%), of whom 1028 were not treated with anticoagulants; subsequent nonfatal VTE occurred in 5 patients (0.5% [95% confidence interval {CI}, 0.2%–1.1%]). Computed tomography showed pulmonary embolism in 674 patients (20.4%). Computed tomography excluded pulmonary embolism in 1505 patients, of whom 1436 patients were not treated with anticoagulants; in these patients the 3-month incidence of VTE was 1.3% (95% CI, 0.7%–2.0%). Pulmonary embolism was considered a possible cause of death in 7 patients after a negative CT scan (0.5% [95% CI, 0.2%–1.0%]). The algorithm was completed and allowed a management decision in 97.9% of patients.

In summary, the Christopher study confirmed that their algorithm to diagnose and treat VTE and PE was safe and efficacious. It should be noted that their societal and clinical practice milieu allowed them to achieve an outstanding recruitment and retention, in addition to a remarkable adherence to the protocol.

4.5.8 Checklist for Observational Studies

- Is this study compliant with the STROBE guidelines? (www.strobe-statement.org)
- Was this a prospective cohort study?
- If it was a case-control study: (A) Is there evidence of Recall Bias? (B) How was the Control group selected? (C) How does the Control group compare to the cases and to your Population?

-
- Who are the participants? How were they recruited? Are they similar enough to your patients?
 - What was the Exposure of interest? Was it appropriately characterized?
 - What was the Outcome of interest? Was it appropriately characterized, in a blinded manner, and independently from the Exposure?
 - Were the follow up completion and duration independent from the Exposure?
 - How was the association between Exposure and Outcome measured? How strong was the association?
 - Was the possibility of Type 1 Error considered? Beware of multiple comparisons...
 - Were there any potential confounders? Were they appropriately measured, and adjusted for?
 - Was there evidence of effect modification (interaction)? If so, was it biologically plausible?
 - Was there a possibility of Bias? You should consider at least the following possibilities:
 - Confounding by Indication;
 - Protopathic Bias;
 - Immortal Time Bias;
 - Attrition of Susceptible Subjects.

Commonly Used Statistics

5

5.1 Relative Risk

When we perform a prospective study, either observational or a clinical trial we can estimate the Relative Risk, which answers the question: How does the exposure *change* the risk of the outcome, relative to the risk at baseline (in the unexposed) over the entire follow-up?

$$\text{Relative Risk (RR)} = \text{Risk in Exposed} / \text{Risk in Unexposed}$$

Suppose you conduct a double-blinded, placebo-controlled, Randomized Clinical Trial (RCT) with 1000 participants in each arm. The exposure in an RCT is the treatment you are evaluating.

After 36 months of follow up, you counted 100 deaths in the intervention arm, and 300 in the placebo (control) arm.

Each occurrence of the outcome of interest is called an event. In our example, each death is an event.

The Relative Risk reflects the change in the risk of events achieved by your treatment. We calculate it as follows:

$$\text{Relative Risk} = \text{Treatment Event Rate} / \text{Control Event Rate}$$

We calculate the event rate (ER) in each group:

$$\text{Event Rate} = \text{Participants with Events in Group} / \text{Group Size}$$

In our example, RR at 36 months = $(100/1000) / (300/1000) = 0.33$.

We like this Relative Risk estimate because of its simplicity. However, it is an accurate reflection of treatment effect only if the three following assumptions are met.

1. There is no significant confounding.
2. Both treatment groups had an equally good follow-up.
3. Differences in risk between the two groups remain similar across the entire follow up period. For example, if one arm fares better than the other, it should do so consistently throughout the entire follow-up.

To see what can be done when the above noted assumptions are not met, please refer to the Hazard Ratio section.

Fortunately, most large RCTs meet all three assumptions. In particular, we can be certain that large RCTs will not suffer from any confounding, because random assignment of large numbers of people will always generate two very similar groups of people, with almost identical distribution of potential confounders.

Like all statistics, Relative Risks are best interpreted when you examine their Confidence Interval as well. Calculating the Confidence Interval of a Relative Risk by hand is too complicated for most of us. However, you can use any online calculator for that. For example, if you go to https://www.medcalc.org/calc/relative_risk.php you obtain the following results (Table 5.1):

The 95% Confidence Interval is narrow, and we are thus confident in our finding of a survival benefit. In addition, the “worst case scenario” in that confidence interval is the upper limit of 0.41, which is still an impressive effect from a clinical perspec-

Table 5.1 Relative risk estimates

Relative risk	0.33
95% CI	0.27–0.41
Significance level	$P < 0.0001$
NNT (benefit)	5
95% CI	4.274 (benefit) to 6.022 (benefit)

CI Confidence interval, *NNT* Number needed to treat

tive. As expected, the P-Value against the Null Hypothesis of no effect on survival is very small (highly significant).

You will notice that the calculator also gave us a “NNT” estimate for the benefit: that is the Number Needed to Treat, which we will discuss after the Relative Risk Reduction.

5.2 Relative Risk Reduction

When you decide to shop online because there is a sale, you probably don't say “I like this sale because all prices are 0.80 of the original value”. Most normal people say: “I like this, because everything is 20% off”. In other words, most people find a relative reduction easier to understand. In the case of our RCT we can estimate the Relative Risk Reduction for mortality, which is done simply by subtracting the Relative Risk from 1:

$$\text{Relative Risk Reduction} = 1 - \text{Relative Risk}$$

In our RCT the RRR was $= 1 - 0.33 = 0.67$.

That means the treatment gave us a 67% reduction in the risk of death.

Not only do most patients understand Relative Risk Reductions quite well, but they also give great value to them. That happens because they tend to care mostly about changes to their individual risk, even if that risk was low to begin with.

On the other hand, to assess therapeutic benefits from the societal perspective we always examine the Number Needed to Treat, which we discuss next.

5.3 Number Needed to Treat

As a society, we are also interested in how the treatment impacts risk in ABSOLUTE terms, because it tells us how many people must be treated to observe a measurable benefit.

For that reason, we always estimate the Number Needed to Treat (NNT).

The Number Needed to Treat is the number of patients who need to be switched from the old treatment to the new one to prevent one bad outcome, over a period of time similar to the study's duration.

We calculate the NNT inverting the Absolute Risk Reduction (ARR):

$$\text{Number Needed to Treat} = 1 / \text{Absolute Risk Reduction}$$

Where Absolute Risk Reduction (ARR) is:

$$\text{ARR} = \text{Control Event Rate} - \text{Treatment Event Rate}$$

Remember: The event rates must be in their decimal point notation.

In our RCT the NNT was:

- $\text{ARR} = 0.3 - 0.1 = 0.2$
- $\text{NNT} = 1/0.2 = 5$

For every five patients you switch to the new treatment, you expect to prevent 1 death over approximately 36 months. This is a remarkable benefit.

For an up-to-date list of published Numbers Needed to Treat, please see <https://www.thennt.com/>

Please note that the NNT estimates benefit over a time period similar to the duration of the corresponding study—in our example, 36 months. For example, if you see the same NNT of 5 but over 12 months, you have a far more efficacious intervention because the same benefit was achieved faster.

The Numbers Needed to Treat may also be estimated from Hazard Ratios and Odds Ratios, but the computation is a bit more complex. To obtain a calculator that does those computations, please go to the calculators section in <https://ebmnotebook.com/>

5.4 Number Needed to Harm

When we examine a negative effect we estimate the Number Needed to Harm.

Instead of inverting the Absolute Risk Reduction (as we do for the NNT) we invert the Absolute Risk Increase.

$$\text{Number Needed to Harm} = 1 / \text{Absolute Risk Increase}$$

Where Absolute Risk Increase (ARI) is:

$$\text{ARI} = \text{Treatment Event Rate} - \text{Control Event Rate}$$

Once again, the event rates must be in their decimal point notation.

5.5 Censoring

Let us say you examine the data from your RCT more closely, and realize that all 700 participants who survived in the control arm had complete follow up for 36 months. On the other hand, of the 900 participants thought to have survived in the intervention arm, 200 had a follow up time of only 24 months—they could not be contacted anymore after that.

How confident are you now about that the Relative Risk of 0.33 is accurate assessment of the treatment effect in this trial?

We said that one of the assumptions underlying our Relative Risk was that both study groups were followed up equally well. In our example, it seems that that assumption was not met. We can use other survival statistics, which apply censoring to solve this problem.

Censoring means that we draw inferences about treatment effect over the entire study aggregating estimates from each one of the time units that compose the follow up. Within each of those time units we calculate Event Rates keeping a subject in the

denominator for as long as her status is known. Once her status becomes unknown, we drop her from the period estimates completely. That means that the denominator for each event rate reflects the completion of follow up for each group at that time. Sounds complicated? Let us examine how Kaplan-Meier curves use censoring.

5.6 Kaplan-Meier Curves

Medical audiences are extremely familiar with Kaplan-Meier curves because they appear in almost every article reporting RCT results. They depict the cumulative probability of survival (or less frequently, the cumulative probability of an event) in each arm of a study.

Let us build a very simple example for just one treatment arm in a small study. We have 100 participants at baseline. Our data show that exactly at 30 days from their enrollment 3 participants died. The Group Probability of surviving at that point was $= 97/100 = 97\%$. The drop in survival probability for the group is easily depicted as the first step down in our curve, as shown in Fig. 5.1.

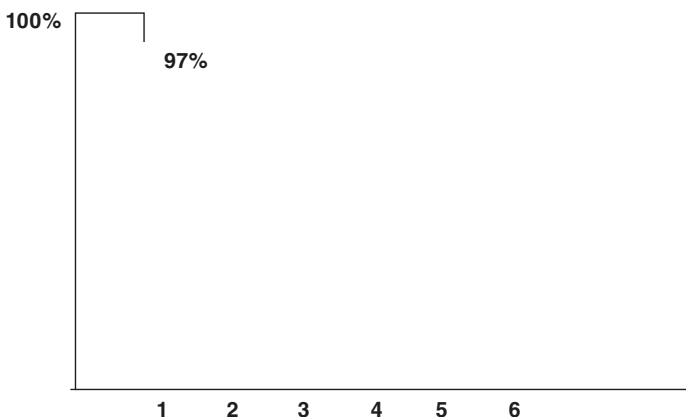


Fig. 5.1 Kaplan-Meier Curve. First drop in survival probability to 97%

We then learn that a fourth participant died exactly 45 days after enrolling. At that point we had lost contact with 2 other participants and we did not know their vital status.

Censoring in this situation means we simply take out those 2 subjects with unknown vital status out of our calculations for the second time period. Therefore, our denominator is now 95, with 1 death from that denominator.

The group probability of surviving to day 45, if you were alive at day 30, is therefore:

Probability of Survival at 45 Days = $94 / 95 = 99\%$.

How do we incorporate this estimate into our curve? Does our estimated cumulative probability of survival go from 97 to 99%, as in Fig. 5.2?

That was obviously not the right approach. However, the solution is simple. It is pretty clear that the probability of surviving the second time period is conditional on having survived the first period. Therefore, the Cumulative Probability of surviving at the end of both of our time periods can be estimated multiplying them, as follows:

Probability of Surviving Periods 1 and 2 = $\text{Prob1} * \text{Prob2}$.

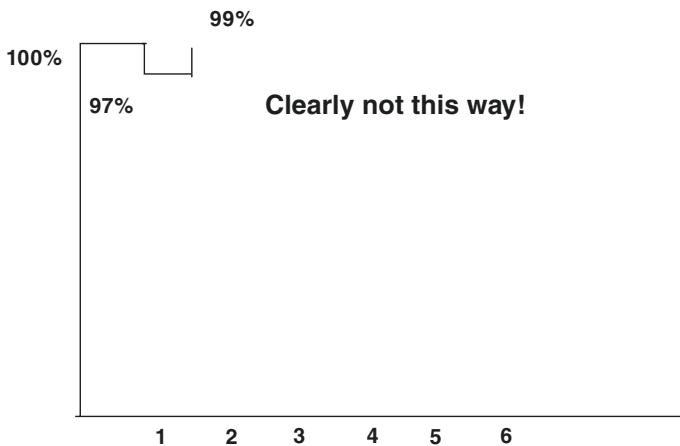


Fig. 5.2 Kaplan-Meier Curve. Effect of using the survival rate from the second set of events without considering the previous one

In our example: Probability of surviving at day 45 = $(0.97 * 0.99) = 96\%$.

That gives us the next step down in the survival curve, as seen in Fig. 5.3. This procedure may be repeated each time we learn of a new death or deaths, censoring those who were not followed up to that point, and multiplying the estimated period probability times the last overall probability.

Our example is just a massive simplification of the Kaplan-Meier method. But the basic concept is the same. Edward L Kaplan and Paul Meier described their method to estimate survival rates when the survival data are incomplete in the June 1958 issue of the Journal of the American Statistical Association. Kaplan and Meier divided the survival curve into discrete time intervals defined by the time of every event. Next, they calculated the mortality risk within each interval as the number of deaths in that interval divided by the number of non-censored patients who were alive at the beginning of the interval. Finally, they calculated a chain of survival rates linking all of the interval event rates by a ‘product-limit estimator’.

Kaplan-Meier estimates are robust but they rely on two main assumptions:

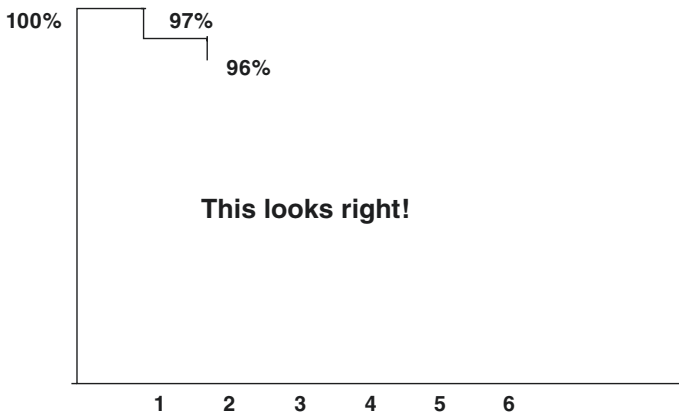


Fig. 5.3 Kaplan-Meier Curve. The second step correctly applies the cumulative approach to estimate survival probabilities

- 1. At any point those patients who are censored have the same survival prospects as those who continue to be followed. In fancier words we say that censoring must be uninformative in regards to the probability of survival.
- 2. We assume that the survival probabilities are the same for subjects recruited early and late in the study.

5.6.1 Incorrect Kaplan-Meier Formatting

A pharmaceutical company mails you a glossy brochure with Fig. 5.4, which depicts the mortality reduction achieved by their new fictional panacea, Wonderil, as compared to placebo in people with mild non-ischemic cardiomyopathy. The Hazard Ratio (95% CI) was 0.61(0.28 to 1.26).

Let us examine in Fig. 5.5 what the graphic looks like if the vertical axis correctly goes from 0 to 1, instead of going from 0.7 to 1. Not that impressive anymore.

This is a reminder that the vertical axis should not be truncated. Rather, it should go from 0 to 100%. A truncated vertical axis

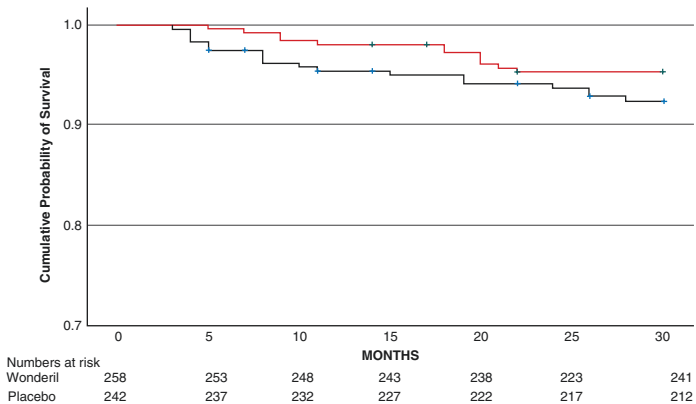


Fig. 5.4 Kaplan-Meier Curves with Truncated Vertical Axis. The vertical axis starts at Relative Risk of 0.70, creating a deceptively strong impression of divergence between the curves

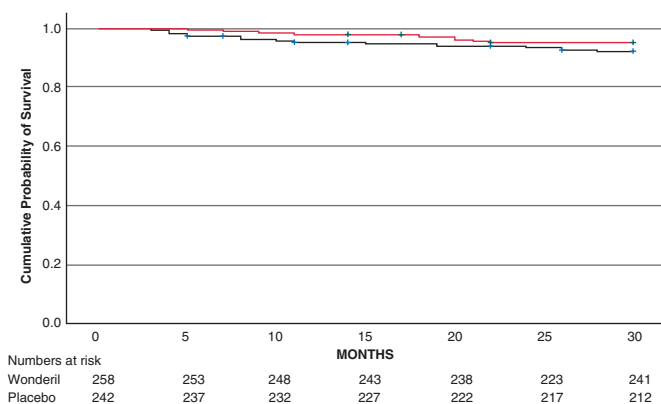


Fig. 5.5 Kaplan-Meier Curves with Appropriate Vertical Axis. The vertical axis now starts at 0, and there is little discernible divergence between the curves

artificially increases the divergence between the two curves, providing an exaggerated visual impression of treatment benefit. A rather cheap trick, but it can be unintentional because the statistical software may generate a truncated graphic automatically.

5.6.2 Censoring and Numbers at Risk

Let us also examine the Numbers at Risk below the horizontal axis. What are they, and how do they help us assess the data depicted in the graphic?

The Number at Risk is the number of subjects that provided that length of follow up data and are, therefore, “at risk” to have an event at that point.

The number at risk inform the confidence we may have on that portion of the curve (lower numbers → less confidence). They are equivalent to having confidence intervals around each curve.

Now, let us consider why the numbers at risk drop as time goes on. Did everybody else die or get lost to follow up? Absolutely not—many of them got censored, because many people got recruited later in the study and did not contribute data up to the end of the curves. Remember that when we create the graphic Time Zero is the recruitment day for all participants, who were actually recruited over a long time.

The overlap between recruitment and follow up can have a huge impact on censoring. In Fig. 5.6 Participant A (blue bar) was recruited on the first day of enrollment, whereas Participant B (red bar) was recruited on the final day of enrollment. They were both alive and they were contacted by the study team at the end of the follow up.

The length of follow up data provided by Participant A was much greater. Therefore, when the curves were generated Participant A was in the “Number at Risk” at the right end of the graphic in Fig. 5.7 (blue bar), whereas Participant B was in the Numbers at Risk for a much shorter time (blue red bar). Participant B was CENSORED at 10 months.

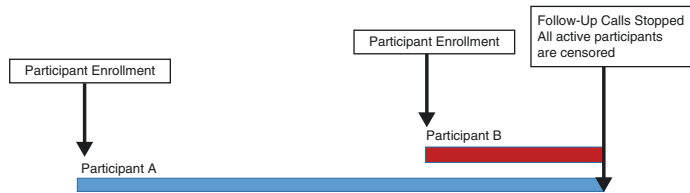


Fig. 5.6 Follow-Up Time Can Vary Greatly. Participant A has a longer follow-up time when study stops calling all participants

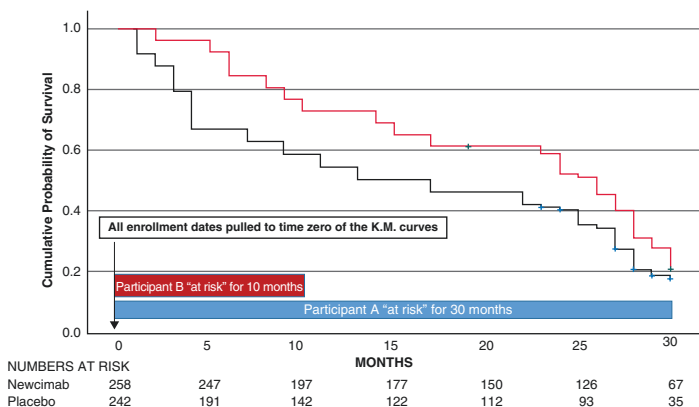


Fig. 5.7 Contribution to Numbers at Risk Depends on Follow-Up Length. Participant B was followed up for 10 months, was censored at that point, and was no longer “at risk” after that

5.7 Hazard Ratio

The most commonly used method to estimate the Hazard Ratio is the Cox Proportional Hazards Model (frequently referred to just as the Cox PH Model), created by Sir David Cox around 1972.

The Cox PH model is a regression method for survival data. It examines treatment effects as differences in hazard rates between the treatment groups, over time.

The hazard rate is the probability that if the event in question has not already occurred, it will occur in the next time interval, divided by the length of that interval. The time interval is made very short, so that in effect the hazard rate represents an instantaneous rate.

The hazard ratio is the ratio of the hazard rate in the treated over the rate in the control group. It is aggregated over all time periods in the study.

$$\text{Hazard Rate} = \text{instantaneous risk in a group}$$

Hazard Ratio = hazard rate in the treatment arm divided
over the hazard rate in the control arm

If a study is quite large and has excellent follow-up completion in both arms, the Unadjusted Hazard Ratio will be very similar, if not identical, to the Relative Risk.

The Hazard Ratio is also similar to the Relative Risk in its interpretation (e.g. a Hazard Ratio = 0.80 is interpreted as 20% reduction in risk per time unit), but it has its own advantages and limitations.

I like to describe the Hazard Ratio as the Relative Risk on steroids, because it does several things well:

- It uses Censoring to adjust for any differences in follow-up length and/or completion.
- It can adjust for Confounding: This is useful in observational studies and in smaller RCT's, particularly if we were not able to stratify the randomization by all the potential confounders.
- It can test for Interactions, which are differences in treatment effect depending on who the patient is (e.g., does my treatment work similarly in people with or without diabetes?) This type of subgroup analysis is usually the final figure of articles reporting RCT results.

5.7.1 Benefits of an Adjusted Hazard Ratio

In observational, non-randomized, studies it is always a good idea to adjust your Hazard Ratio estimate. Adjusted Hazard ratios are also useful in small Randomized Clinical Trials (RCT). In a small RCT, you may not be able to stratify your randomization for every possible confounder, leaving you no choice but to adjust your analysis for the most important confounder.

For example, the TOPPS study randomly assigned patients undergoing chemotherapy or stem-cell transplantation to either receive, or not to receive, prophylactic platelet transfusions when

morning platelet counts were less than 10×10^9 per liter (*N Engl J Med* 2013; 368:1771-1780). The primary end-point was severe bleeding up to 30 days after randomization. It had a relatively small size of 598 participants, creating concerns for imbalances in randomization and consequent confounding. For that reason, the Cox Proportional Hazards model was adjusted for the type of malignancy, and for the type of oncological treatment received (chemotherapy, stem cell transplantation, etc.).

Of the 301 participants in the control arm 151 had a major bleeding, whereas that happened in 128 of 299 participants in the intervention arm. The adjusted Hazard Ratio for major bleeding without prophylactic platelet transfusion was 1.30 (1.04–0.1.64).

If we estimate the ‘crude’ unadjusted Relative Risk, we obtain:

$$\text{Relative Risk} = (151 / 301) / (128 / 299) = 1.17 (0.99 - 1.39)$$

Clearly, the unadjusted Relative Risk is less impressive than the adjusted Hazard Ratio, and its confidence interval crosses 1, suggesting lack of statistical significance. This example shows how the adjusted Hazard Ratio can provide greater statistical power to detect an effect than the unadjusted Relative Risk.

5.7.2 Assessing Palliative Treatments

When median survival improves but the overall long-term mortality remains high, the Hazard Ratio also tends to capture the treatment benefit better than the Relative Risk. This is most evident in RCTs of palliative cancer treatments.

Let us examine an example of a palliative treatment. Our example compares survival with a fictional new drug, Newcimab, to standard of care chemotherapy.

Many participants had events at the end of 30 months, but there was a significant difference in median survival between the 2 groups. The Cox Hazard Ratio was 0.72 (0.59 to 0.88). The median survival time on Placebo was 12.1 months, whereas on Newcimab it was 24.6 months.

The separation between the two Kaplan-Meier curves for most of the follow-up is evident in Fig. 5.8. However, we may not appreciate the improvement achieved by Newcimab if we limit our comparison to a crude Relative Risk at 30 months as shown next (Table 5.2).

The Hazard Ratio was more sensitive to detect the improvement in median survival times than the crude Relative risk. That was also shown by the divergence between the Kaplan-Meier curves, in spite of the fact that they converged towards the end.

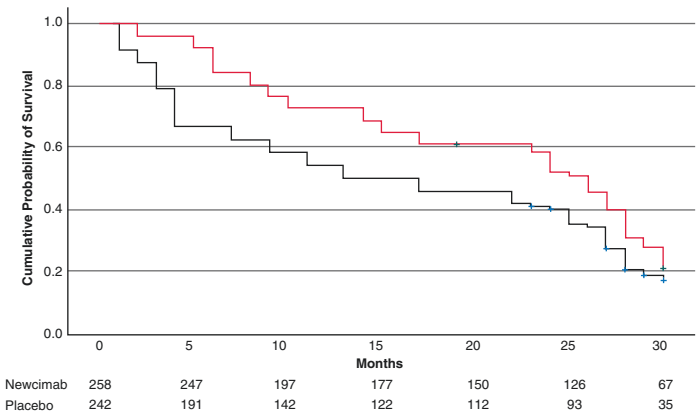


Fig. 5.8 Kaplan-Meier Curves with Late Convergence. There is clear divergence for most of the follow-up, but the curves converge later

Table 5.2 Relative risk based on cumulative event rates

Number of events in the control group:	196
Sample size of the control group:	242
Number of events in the experimental group:	198
Sample size of the experimental group:	258
Relative risk:	0.947
CI:	0.865 to 1.037

5.7.3 The Proportionality Assumption

The main assumption underlying the Cox method is that once the survival trends in each arm diverge from each other, they will always do it in the same direction. In other words, if you plot them as survival curves the curves should not cross each other over time, as they do in Fig. 5.9.

That figure shows a clear violation of the proportionality assumption. When this happens, the Hazards should be estimated using a method different from the Cox PH model.

The presence of Proportionality is also required when you perform another popular survival test, the Log Rank statistic.

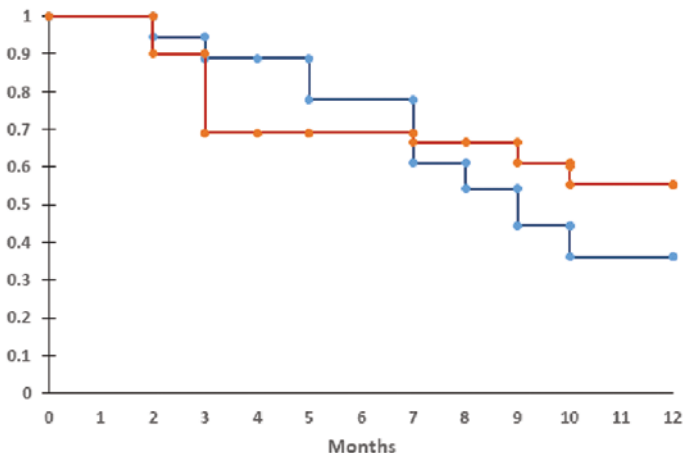


Fig. 5.9 Lack of Proportionality. The Kaplan-Meier curves cross each other, because the risks do not remain proportional over time

5.8 The Log-Rank Statistic

The log-rank test is very useful to examine differences in time to event between treatment groups. For that reason, you will frequently see it used in Randomized Clinical Trials (RCT). The Kaplan-Meier survival curves, a practically mandatory component in RCT reports, commonly include either in the graphic itself or in the footnote a P-Value from a log-rank test.

The log-rank tests the null hypothesis that there is no difference between the study arms in regards to the probability of an event at any time point. The analysis is performed at the time of each new event. At each such time, the log-rank calculates the observed number of events in each group and the number of expected events if there were no difference between the groups. Let us examine an example using a very simple summarized data set.

Let us say we have an RCT with two treatment arms, control and intervention, each with 100 participants. They suffer from a condition with a high risk of death within a couple of months. We have summary follow up data for 1-month periods, over 3 months.

During those 3 months there were 30 deaths and 2 participants were lost to follow-up, as seen in Table 5.3.

The log-rank compares the proportion of observed to expected events between the 2 arms. For every month, we will build a 2×2 table that represents our data (Observed Events), and contrasts that to the Expected Events if the Null Hypothesis (H_0) were true. If the H_0 were true we would have identical event rates in both arms.

Table 5.3 Follow-up data over 3 months

	Control (100)	Intervention (100)
1	4 deaths	2 deaths
2	8 deaths	0 deaths (1 censored)
3	12 deaths (1 censored)	4 deaths

We create a table for each time period, and all the resulting tables are aggregated into a single final table, which can be assessed with a simple chi-square test.

5.8.1 Observed Vs. Expected Event Rates in Month 1

During the first month we had 6 deaths overall, in 200 subjects. There were 4 deaths in the control arm and 2 deaths in the intervention arm.

We can estimate how many events we would EXPECT to see in each arm if the event rates were identical (if the null hypothesis of no difference were true).

Thus, the expected overall event rate was $6/200$.

We calculate the expected number of events in an arm with n subjects multiplying the expected event rate times the sample size of that arm:

$$\text{Expected Events} = n * (6 / 200)$$

We expected to see $[n * (6/200)]$ deaths in each arm. Because both arms had 100 subjects, the expectation is: $(6/200) * 100 = 3$, in both. Our summary 2×2 data for month 1 is shown in Table 5.4.

5.8.2 Observed Vs. Expected Event Rates in Month 2

During the second month we had 8 deaths overall, all of which took place in the control arm. We do not know the status of 1 participant in in the intervention arm. We take that participant out of

Table 5.4 Observed and Expected Events in Month 1

	Control ($n = 100$)	Intervention ($n = 100$)
Observed	4	2
Expected	$6/200 \times 100 = 3$	$6/200 \times 100 = 3$

the month’s denominator for that arm and from the overall sample size: we censor them.

Once again, we estimate how many events we expect to see in each arm in the second month if the event rates were identical (under the null hypothesis scenario). During the second month we had 8 deaths overall, in 193 subjects (200 minus the 6 dead, and 1 censored). Once again, we calculate the expected number of events in an arm with n subjects multiplying the expected event rate times the sample size of that arm:

$$\text{Expected Events} = n * (8 / 193)$$

Our summary 2x2 data for month 2 is shown in Table 5.5.

5.8.3 Observed Vs. Expected Event Rates in Month 3

During the third month we had 16 deaths overall, in 184 subjects (200 minus the 14 dead, and 2 censored).

$$\text{Expected Events per Arm} = n * (16 / 184)$$

The summary data for Month 3 is shown in Table 5.6.

Table 5.5 Observed and Expected Events in Month 2

	Control ($n = 96$)	Intervention ($n = 97$)
Observed	8	0
Expected	$8/193 \times 96 = 3.97$	$8/193 \times 97 = 4.02$

Table 5.6 Observed and expected events in month 3

	Control (87)	Intervention (97)
Observed	12	4
Expected	$16/184 \times 87 = 3.89$	$16/184 \times 97 = 8.43$

5.8.4 Obtain a Global Chi-Square Test

We add up all cells, creating a single 2x2 table (Table 5.7) with rounded numbers (10.86 becomes 11, and 15.45 becomes 15).

We analyze the table with a chi-square test with one degree of freedom (see below). Our P-Value is highly significant. This is the simplest example of a log-rank test.

$$\text{Chi-Square} = 8.43 \quad \text{degrees of freedom} = 1 \quad \text{P-Value} = 0.003$$

Please note that our example is just that, an example. When we perform log rank tests using software we estimate observed vs. expected events every time there is a new event—not once a month.

Just like the Kaplan-Meier curves, the log rank test assumes that those censored have the same survival prospects as those not censored—in other words, censoring is uninformative. We also assume that risk of an event remains similar, regardless of whether it is early or late in the study. If those assumptions are not met, other methods should be used.

In addition, the log-rank is only a significance test and it does not provide an effect size estimate or a confidence interval.

Table 5.7 Cumulative estimate of total number of events

	Control	Intervention
Observed	24	6
Expected	10.86	15.45

5.9 Odds Ratio

Let us now briefly review the odds ratio, which is mainly used when better statistics, such as the relative risk, may not be applied. That is the case for cross-sectional and case-control studies, in which a prospective ascertainment of incident disease is impossible, preventing us from relative risks.

In its simplest form the univariate odds ratio evaluates the significance of a single exposure for a given outcome. It can be summarized as the “cross-product ratio” from a 2x2 table (Table 5.8).

The cross-product ratio has the product of the “true cells” in the numerator and the product of the “false cells” in the denominator.

We define as “true” the cells consistent with the hypothesis of association, whereas we call “false” the cells consistent with the null hypothesis of no association.

$$\text{Odds Ratio} = (a * d) / (b * c)$$

Of note, when there are multiple exposures to assess we may obtain multivariate adjusted odds ratios using multiple logistic regression, as we discussed when we reviewed propensity scores in the Observational Studies module.

Regardless of whether they were obtained using univariate or multivariate methods it is important to remember that inflated odds ratios may be seen when the outcome prevalence is high. This very well-known phenomenon is sometimes forgotten.

A useful rule of thumb is to suspect odds ratio inflation whenever the outcome prevalence is greater than 10%. The greater the prevalence, the higher the risk of odds ratio inflation.

Table 5.8 Two-by-Two Table for Outcome by Exposure

	Outcome	No outcome
Exposure	a	c
No exposure	b	d

5.10 Vaccine Efficacy

Vaccine efficacy is usually estimated as the reduction in disease risk (incidence after vaccination) from the baseline risk (incidence in those unvaccinated), relative to the baseline risk.

Vaccine Efficacy:
 $(\text{Risk in Unvaccinated} - \text{Risk in Vaccinated}) / \text{Risk in Unvaccinated}$

Of note, we should start to count incident cases after enough time has passed to allow an immune response from the vaccine. For the COVID-19 vaccines that time was 2 weeks.

The immune response to a vaccine may wane during longer term follow-up. To characterize that phenomenon, we can measure periodically age-adjusted incidence rate ratios (IRR) for cases among unvaccinated persons compared with those cases among vaccinated persons. Age adjustment is a statistical process that applies weights to the rates of health outcomes, allowing communities with different age distributions to be compared. To see an example of how IRRs are used please see the CDC publication about COVID-19 vaccines authored by Johnson and colleagues [Johnson AG, et al. *COVID-19 Incidence and Death Rates Among Unvaccinated and Fully Vaccinated Adults with and Without Booster Doses During Periods of Delta and Omicron Variant Emergence — 25 U.S. Jurisdictions, April 4–December 25, 2021. MMWR Morb Mortal Wkly Rep* 2022;71:132–138. <https://doi.org/10.15585/mmwr.mm7104e2>].

When full vaccination requires more than one dose, as is the case for several COVID-19 vaccines, we must estimate efficacy for those fully vaccinated separately from those who have received only one vaccine dose.

5.11 Attributable Proportion

The Attributable Proportion statistic is used in Public Health to measure the effect of a causal factor on the risk of an outcome. Appropriate use of this statistic is based on three assumptions: (1) the causal factor increases the risk of the outcome; (2) the risk of the outcome in those not exposed to the causal factor represents the baseline risk; (3) the risk factor of interest is solely responsible for the excess risk. It is estimated as follows:

Attributable Proportion:

$(\text{Risk for Exposed} - \text{Risk for Unexposed}) / \text{Risk for Exposed}$



Randomized Clinical Trials

6

6.1 Why Do We Need Randomized Trials?

Observational studies suffer from biases and confounding, which can never be completely eliminated. Whenever feasible and ethically appropriate, a randomized clinical trial is the best means to determine whether a treatment works as expected. This is true no matter how “obvious” or “logical” the rationale for a given treatment might seem.

There have been clear examples of widely accepted and “logical” treatment strategies, seemingly based on a sound understanding of the pathophysiology, but which were revealed as non-efficacious, and sometimes downright harmful, when tested through a well-designed randomized trial.

A notorious and dramatic example of how a randomized trial can prove us wrong was the Cardiac Arrhythmia Suppression Trial (CAST; *N Engl J Med* 1991; 324:781–788). Patients with poor left ventricular ejection fraction after a Myocardial Infarction (MI) are at very high risk of arrhythmic death. For that reason,

those patients frequently received antiarrhythmics when their monitoring detected frequent ventricular ectopy after the MI. We did that with the best possible intentions, and we thought that it made complete sense—until CAST was carried out.

CAST randomized 1500 post-MI patients with increased ventricular ectopy and depressed LV function to either encainide, flecainide or placebo. The trial was stopped early. Treatment with encainide or flecainide was associated with almost three-times higher risk of arrhythmic deaths or cardiac arrests (5.7% vs. 2.2%) with a Number Needed to Harm (NNH) of 29 over a mean follow-up period of only 10 months.

6.2 Types of Clinical Trials by Phase

Phase I: Researchers test a new drug or treatment in a small group of people for the first time to evaluate its safety, determine a safe dosage range, and identify side effects.

Phase II: The drug or treatment is given to a larger group of people to see if it is effective and to further evaluate its safety.

Phase III: The drug or treatment is given to large groups of people to confirm its effectiveness, monitor side effects, compare it to commonly used treatments, and collect information that will allow the drug or treatment to be used safely in the population at large.

Phase IV: Studies are done after the drug or treatment has been marketed to gather information on the drug's effect in various populations and any side effects that are either relatively rare or associated with longer-term use.

6.3 Clinical Trial Registration and Compliance with Guidelines

In order to ensure transparency and validity of clinical trials every trial must be registered in a reputable registry, such as clinicaltrials.gov.

als.org, before any participants are recruited. This type of registry ensures that study designs are prospectively characterized in detail, and they allow monitoring to reduce the risk of intentional publication bias.

That registration is the initial step in a sequence that ends with appropriate compliance with the guidelines that govern the reporting of clinical trials, the CONSORT guidelines. [www.consort-statement.org] CONSORT provides a minimum set of rules aimed at maximizing the quality of reporting of clinical trials. Active implementation of the CONSORT guidelines has been shown by Hopewell and colleagues to improve the quality of abstracts in major medical journals. [Hopewell S, Ravaud P, Baron G, Boutron I. Effect of editors' implementation of CONSORT guidelines on the reporting of abstracts in high impact medical journals: interrupted time series analysis. *BMJ*. 2012 Jun 22;344:e4178. <https://doi.org/10.1136/bmj.e4178>. PMID: 22730543; PMCID: PMC3382226.]

6.4 Inclusion and Exclusion Criteria

A careful definition of our study population is the first essential step in the design of a Randomized Clinical Trial (RCT). It is extremely important because a miscalculation at this step may result in a failed RCT.

The main goal of the Inclusion Criteria is to define the patient population that is likely to benefit from the new treatment.

On the other hand, Exclusion Criteria usually include at least the following:

- Patients who are already known to benefit from the new treatment (it would be unethical to randomize them).
- Patients who have an underlying or superimposed pathophysiology that differs from the one we are targeting.
- Patients with a natural history of the disease that differs from what we expect to see most commonly.

6.5 Internal and External Validity

Overall, we try to recruit participants who will benefit in a predictable manner from the new treatment. Therefore, one of the goals of the exclusion criteria is to “clean up” our experiment. In doing so, we are increasing the **internal validity** of our study.

However, study results will only apply to patients similar to our study sample. Therefore, each exclusion criterion decreases the applicability, or **external validity**, of our findings.

It is clinically important to strike a reasonable balance between internal and external validity when defining inclusion and exclusion criteria. Quite often, that balance is achieved over time. Clinical trials of novel interventions usually start in rather restricted populations, and subsequent studies expand the indication to other groups, but only after initial success has been achieved.

If a clinical trial enrolls a vulnerable population, such as pregnant women, there needs to be a clear assessment of benefits that overwhelming outweigh the potential risks. On the other hand, if we systematically exclude vulnerable populations we may find ourselves doubting whether our research results are fully applicable to them.

6.6 Type 1 and Type 2 Statistical Errors

In RCTs, as in any other study, we do our best to reduce the probability of statistical errors.

Type 1 Error (α) is the probability of reporting a FALSE POSITIVE finding, rejecting the null hypothesis when it is actually true.

We try to keep this probability at less than 5%. To keep the type 1 error at this level, we usually limit our main analysis to

just one, clinically relevant, pre-specified PRIMARY OUTCOME. All other analyses are secondary. We do that because every additional analysis increases the overall probability of a false positive finding.

Type 2 Error (β) is the probability of reporting a FALSE NEGATIVE finding, accepting the null hypothesis when it is actually false.

We try to keep this probability at 20% or less. Power is our ability to detect a treatment effect when it is real, thus appropriately rejecting a false null hypothesis. Power is $1 - \beta$. Therefore, we wish to have a power of at least 80%. The most important determinant of Power is our Sample Size.

6.7 Sample Size and Power Estimates

To ensure that we have adequate Statistical Power it is crucial to estimate accurately how many participants our study needs. By convention, we estimate our sample size aiming for a Power of at least 80%, and limiting our main analysis to one Primary Outcome, which is significant if the P -Value is <0.05 . The other items considered to estimate the Sample Size are:

- Expected event rate in the control group. For example, what is the expected mortality for people on placebo, or receiving an older drug? The higher the baseline risk, the easier it is to achieve significance with a smaller sample. In other words, a lower event rate in the control arm results in the need for a larger sample size to achieve significance when trying to prove that the new treatment is better.
- Expected benefit from the new treatment (e.g., mortality reduction achieved by the new drug). A less efficacious treatment requires a larger sample size to achieve statistical significance.

For that reason, an excessively optimistic prediction of benefit is frequently the cause of insufficiently powered studies.

- Expected attrition/losses to follow up. If we expect that many patients may be lost to follow up, we should compensate for those losses upfront, recruiting a larger number of participants.
- Expected crossovers. This is particularly important when the intervention cannot be blinded, and many patients randomized to the control arm may actively seek the new treatment.

All these items must be carefully considered when deciding the sample size, lest we end up with insufficient statistical power.

6.8 Randomization: The R in RCT

We randomly assign the treatment to avoid BIAS and to reduce the probability of CONFOUNDING.

6.8.1 Preventing Bias

We avoid bias because we do not allow investigators to select who will receive each treatment. Neither do we allow participants to elect which arm they prefer to be in. For example, if investigators were allowed to preferentially assign sicker participants to the control arm, the intervention might appear to perform better just because of that biased treatment assignment.

Randomization should be performed in an automated, blinded manner, using remotely accessible software, which randomizes participants and registers that randomization status in a non-modifiable database. This process keeps investigators blinded from the very beginning, and it is known as Allocation Concealment.

In a well-designed RCT we preserve the effects of randomization and stay away from bias through the entire study when we do the following:

- **Blinding of treatment (double blinding, if feasible).**
- **Complete follow-up of both groups by study personnel who are also blinded to the treatment assignment.**
- **Review and ascertainment of the events by blinded researchers.**
- **The analysis of results is performed applying the Intention to Treat Principle. Intention to Treat means that we analyze participants in the arm to which they were randomized, regardless of what treatment they actually received. A more intuitive name of this approach is “Analyze as Randomized”.**

6.8.2 Reducing Confounding

In addition to avoiding BIAS, randomization also reduces the probability of CONFOUNDING.

In fact, in very large RCTs (mega-trials with sample sizes in the thousands) randomization almost invariably prevents confounding, because the random assignment of a large number of people ensures that almost every important characteristic is equally distributed across the study groups. The only exception may be the imbalances that may take place within study sites.

Imagine you have ten different study sites. One of them is an exceptionally good academic medical center whose patients are

affluent and have no socioeconomic barriers to care. If we get an unevenly higher distribution of participants to the intervention arm in that site, we may create confounding by site because those participants may fare better because of their local characteristics rather than study treatment. We prevent this type of confounding stratifying the randomization by recruitment site.

On the other hand, smaller studies must always prevent confounding, ensuring that the distribution of potential confounders is even across study arms at the end of recruitment. We can achieve that goal using stratified randomization and blocked randomization.

6.8.3 Stratified Randomization

To avoid confounding we frequently stratify the randomization based on expected confounders. In addition, to prevent confounding by differences between study sites, multicenter studies commonly stratify their randomization by site. For example, if we thought that the response to our intervention may be confounded by the presence of myocardial ischemia, and we were also concerned by the possibility of confounding by site, we may stratify our randomization process as depicted in Fig. 6.1. Each bin in this figure represents the pool of random numbers from which the statistical software draws the treatment assignment, stratifying by site and by ischemic status.

6.8.4 Randomization by Blocks

Imagine we wanted to randomize participants on a 1:1 ratio, but as we go along there is an imbalance that deviates from the desired ratio. Blocks are used to prevent that kind of problem.

The use of blocks ensures that as we proceed we obtain a balanced number assigned to each arm within the block, at each study site. For example, randomization in blocks of 10 can ensure that for every 10 participants we randomize, 5 will go to each arm.

In Fig. 6.2 we find the first randomization block for each of four study sites in a New York City RCT. When each center ran-

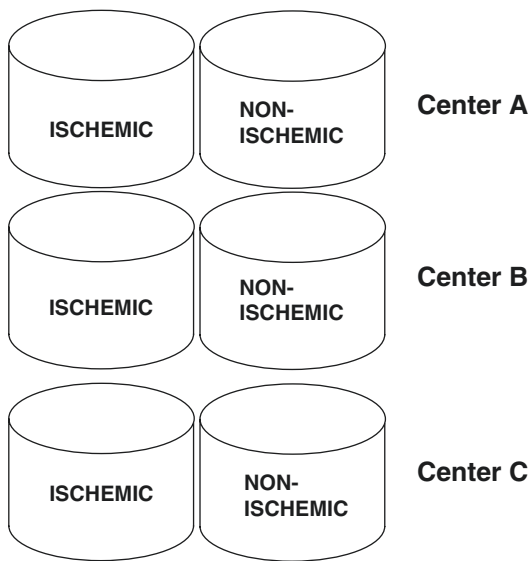


Fig. 6.1 Center-Specific Stratification by Ischemic Status. Participants are randomized stratified by center, and within center, by ischemic status

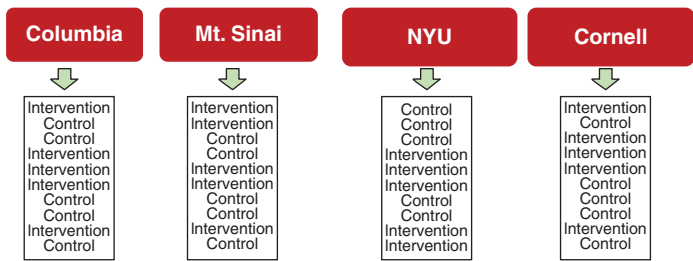


Fig. 6.2 Randomization by Permuted Blocks. The order in which participants are randomized to one of the two arms is different (permuted) in each block

domizes a block of 10 participants, 5 will have gone to each arm. Compare the order within each ach block and you will see that each block is unique, it is permuted, from block to block; and that is why we call them Permuted Blocks.

6.8.5 Did Randomization Prevent Confounding?

To assess whether randomization worked in preventing confounding in a study we are reading, we should examine the table that describes Baseline Characteristics by treatment arm (it is usually the first table in most RCT reports).

It should be noted that our assessment must be based on clinical judgement: we need to assess whether the difference seems large enough to result in an impact on observed outcomes. If we believe there may be confounding, we should consider adding a sensitivity analysis that adjusts for the putative confounder.

It is not a good idea to use *P*-Values obtained from comparisons across post-randomization groups to determine whether there might be confounding due to imbalanced randomization. *P*-Values can be rather misleading in this context, for the following reasons:

1. If you have a mega-trial (with thousands of participants per arm) a tiny, clinically trivial difference between study arms can give you a “significant” *P*-Value <0.05 , even though it will not be a likely cause of confounding.
2. Comparing one group to another numerous times increases the probability of a random, false positive, finding of “significant” difference.
3. In a small study, when confounding is a real risk, *P*-Values may be >0.05 even when the post-randomization imbalance seems clinically relevant.
4. And last, but certainly not least, the epistemological issue: Null hypothesis significance tests were not created to compare two groups created through randomization from the same pool of participants.

6.8.6 Clustered Randomization

Clustered Randomized Clinical Trials are the best option when we expect participants in a study, their environment, and/or their doctors to:

- (A) Influence each other regarding their response to treatment, behavior.
and/or
- (B) Behave or respond similarly, due their shared environment.

A clustered design starts at randomization: you randomize the entire cluster to a study arm. A cluster may be each hospital, each intensive care unit (ICU), or a group of similar hospitals or ICUs.

For example, Huang et al. [*N Engl J Med* 2013; 368:2255-2265] compared three methods to attempt Methicillin-Resistant *Staphylococcus Aureus* (MRSA) decolonization in ICUs. They studied three methods: targeted decolonization, universal decolonization, or screening and isolation. They used a very elegant randomization approach that combine stratification with clustering. They used stratification to optimize balance in patient volume and baseline prevalence of MRSA carriage on the basis of clinical cultures and screening tests done before the trial. Hospitals then were ranked according to ICU volume and were grouped into sets of six. Within each set, they ordered the hospitals according to the prevalence of MRSA carriage in the ICU. Each group of three consecutive hospitals was randomly assigned, one to each strategy group, with the use of block randomization.

Their clustered design allowed them to compare event rates before and after randomization for each treatment group, adjusting for site-specific risk. They ran proportional-hazards models with shared frailties that accounted for clustering within each hospital. They showed that universal ICU decolonization was more effective.

The methodology used by Huang et al. exemplifies how the best statistical analysis takes into consideration the clustering created at randomization.

Alternatively, you may also encounter studies that use mixed models as the analysis method. Mixed models are state-of-the-art statistical tools that can adjust for clustering effects and for participant-level differences, both at baseline and over time.

6.8.7 Fixed Unequal Allocation

The simplest approach is always to randomize participants in equal proportions of 1:1. However, sometimes researchers use Unequal Allocation, and randomize participants in other ratios, such as 2:1 or 3:2.

There are several reasons why unequal allocation can be appealing:

1. A larger sample size in the novel treatment arm gives us greater statistical power to characterize the risk of adverse events. For example, if one of the study arms receives a novel therapy that still needs evaluation data for FDA approval, we may randomize more participants to receive that therapy, allowing us to better describe its safety to the FDA.
2. When we design the study to allow more participants to gain access to a new treatment, we can expect recruitment to be enhanced. For example, cancer patients may be enrolling because they want access to a new experimental treatment. We can design our randomization so they are more likely to receive the new treatment. Of note, this is also ethical if we have good reasons to believe the new treatment is substantially better.

Of note, any deviation from 1:1 randomization will reduce statistical power and will thus require a larger sample size.

6.8.8 Dynamic Allocation: Adaptive Randomization and Minimization

In smaller clinical trials, the use of stratification and blocks are good means to achieve a balanced distribution of covariates across randomized groups if the potential confounders are not numerous and our sample size is large enough to “fill in” the strata/blocks as we proceed. However, those methods may be unable to cope with the complexities of smaller studies that recruit very sick participants with many potentially important confounders, as is often the

case in studies in intensive care settings. An algorithm for adaptive randomization can be created to deal with that type of challenge.

There are many different methods that allow us to adapt and deviate from a 1:1 randomization schedule as needed, but nowadays they all use software to detect imbalances and correct them. They are known as adaptive randomization and minimization. They used to be described as “complicated” in the literature, a rather quaint characterization in our current era of flexible software coding.

Let us examine an example. Imagine we have a new antiviral that substantially reduces the clinical progression of COVID-19 infection, reducing hypoxemia and the need for oxygen supplementation, as compared to an older antiviral. Our budget limits our expected sample to only 400 participants in each study arm. How do we ensure a balanced distribution of covariates across our study arms?

We start by identifying the clinically important covariates and defining how strong a confounder they might be. That will inform our programming in regards to how big an imbalance should be tolerated. For example, C-reactive protein (CRP) levels, age, gender, and PaO₂ are highly predictive of poor outcomes in COVID-19, and are thus a potentially strong confounder. We can use all those covariates to create a minimization score. Our randomization starts at 1:1 but we monitor all pre-defined covariates and the resulting minimization score in each arm as we go along. For example, after randomizing ten participants the program detects that the minimization score is significantly higher in one of the study arms. The eleventh participant will be assigned in a way that provides a better balance in minimization scores across the two groups.

Of note, the treatment assignment of the eleventh participant in our example was non-random. However, that does not reduce the validity of our overall randomization scheme for several reasons:

1. All ten preceding participants were assigned randomly, which makes the assignment of the eleventh participant quasi-random.
2. The entire procedure is automated, pre-defined, and blinded to researchers (i.e., it preserves allocation concealment).

6.9 Blinding

To preserve the benefits of randomization and stay away from bias, we always try to implement double blinding. This means that both participant and investigator are blinded to the treatment assignment.

Lack of blinding may cause bias in a randomized trial. Performance bias is the specific instance in which participants seek additional care or exhibit healthier behavior when they know they are receiving the new treatment, or when clinicians provide better care to them because of the same reason.

Sometimes double blinding is not feasible, as when we compare a surgery versus medical treatment. In that situation, we must ensure that all the components of the outcome evaluation remain blinded as follows. The personnel performing the follow-up, the researchers determining whether the outcome took place (ascertaining the outcome), and those analyzing the data, should all be blinded to the treatment assignment.

In addition, if double blinding is not feasible, we must avoid using “softer” outcomes, such as quality of life scores, because the participants’ report may be biased if they know what treatment they are receiving.

Let us examine an example of the possible consequences of carrying out an un-blinded clinical trial.

Renal artery denervation is an endovascular nerve ablation procedure, which has been used for the treatment of resistant hypertension (blood pressure $\geq 140/90$ mm Hg despite the use of ≥ 3 antihypertensive medications). The Symplicity HTN-2 Trial was an un-blinded randomized controlled trial that compared renal denervation to usual care. It achieved remarkable success (*Lancet* 2010; 376: 1903–09). Between-group differences in blood pressure at 6 months were 33/11 mm Hg for systolic and diastolic blood pressures, respectively ($P < 0.0001$ for both). At 6 months, 41 (84%) of 49 patients who underwent renal denervation had a reduction in systolic blood pressure of 10 mm Hg or more, compared with 18 (35%) of 51 controls ($P < 0.0001$).

However, when a double blinded design, with a sham procedure, was used in the Symplicity HTN-3 study the results were quite different (*N Engl J Med* 2014; 370:1393–1401). That study did not show a significant reduction of systolic blood pressure in patients with resistant hypertension 6 months after renal-artery denervation, as compared to the sham control group. Both groups had better blood pressure control at 6 months, when compared to their own baseline.

Of course, certain interventions cannot be blinded and use a sham procedure. A good example of that are the studies of Implantable Cardiac Defibrillators (ICD). MADIT II was one of those studies (*Prophylactic Implantation of a Defibrillator in Patients with Myocardial Infarction and Reduced Ejection Fraction, N Engl J Med* 2002; 346:877–883). MADIT II assessed mortality reduction through prophylactic implantation of ICDs in patients with reduced left ventricular function after a myocardial infarction. They assigned participants in a 3:2 ratio to receive an implantable defibrillator or conventional medical therapy. Treatment assignment could not be blinded.

However, we do not worry about a biased result in MADIT II because they had mortality as the outcome of interest, and they have excellent completion of follow up. Only three participants were lost to follow up by the end of the trial, and they were all known to be alive within 6 months before the trial ended.

6.9.1 Involuntary Unmasking

In the AHeFT study 1050 African-American patients who had New York Heart Association class III or IV heart failure with dilated ventricles were randomly assigned to receive a fixed dose of isosorbide dinitrate plus hydralazine or placebo in addition to usual care for heart failure (*N Engl J Med* 2004; 351:2049–2057). The primary end point was a composite score made up of weighted values for death from any cause, a first hospitalization for heart failure, and change in the quality of life. The change in quality of life was assessed with a questionnaire.

The use of a subjective quality of life component in the primary outcome was problematic. Consent Forms describe both adverse effects and potential benefits in detail. If participants are able to guess whether they are taking the active treatment, they may be unconsciously biased to feel an improvement in their disease specific symptoms. In AHeFT those receiving the active intervention did experience headache and dizziness more frequently than placebo patients, so they may have been predisposed to report improvement in their heart failure symptoms, too.

6.10 Crossovers

A crossover takes place when a participant randomly assigned to a treatment arm ends up receiving the other treatment.

Because all good RCTs perform Intention to Treat Analysis, crossovers will result in a loss of statistical power. In other words, crossovers increase the probability of Type 2 Statistical Error.

It should be noted that the direction in which participants cross over (from intervention to control or from control to intervention) does not matter.

All crossovers reduce the statistical power of the study, regardless of their direction.

When researchers expect crossovers during their study they should compensate for that by recruiting more participants. In other words, expected crossovers should go into the sample size estimate to ensure adequate power.

6.11 Completeness of Follow-Up

A good RCT needs a complete follow-up in both arms of the study.

The worst-case scenario is when participants in the intervention arm are lost to follow-up more frequently than those in the control arm. This could be an indication of serious side effects from the intervention, which make participants unable and/or unwilling to engage with the follow-up.

If follow-up is incomplete, we must assess whether there has been Informative Censoring. When we lose a participant to follow-up, we censor them at that point. Informative censoring takes place when censored subjects are either more or less likely to experience the event than remaining individuals in the future. Let us examine an example. We compare the effect on 2-year survival of two treatments (A and B) for non-small cell lung cancer. At the end of the 2 years, we have lost to follow-up about 30% of participants in each arm. For treatment A, the dropouts were mostly due to treatment failure, whereas for treatment B dropouts happened when participants were cured and decided to move to the Caribbean. Disease-free survival rates would be overestimated for intervention A and underestimated for intervention B, because they would both be based on the patients who stayed in the study. We diagnose Informative Censoring if we can show that the same factors associated with poor outcomes early in the follow-up can predict subsequent dropouts in study arm A, but not in study arm B. A similar approach examines whether censoring is associated to risk predictors that were already well characterized before we conducted our study.

6.12 Intention to Treat Analysis

With very few exceptions, we always analyze participants within the arm they were randomized to, regardless of whether they crossed over or stopped taking the study drug. That is described as the Intention to Treat principle. A better name for it is Analyze as Randomized. We do this to preserve the effects of randomization and prevent bias and manipulation.

We mentioned exceptions. Those exceptions are Modified Intention to Treat and As Treated analyses.

We use a Modified Intention to Treat approach to adapt to real life situations while staying away from bias. Let us examine an example. You are evaluating the efficacy of Pre-Exposure Prophylaxis (PREP) to prevent HIV seroconversion of HIV-negative patients. You are enrolling participants from lower income areas of Sao Paulo, Brazil. All participants undergo screening with HIV tests and, if eligible, are randomized to PREP or placebo in a double-blinded design. However, it is complicated to bring participants into the clinics to give them the study drugs, and the average delay between randomization and that clinic visit is 4 weeks. For that reason, you obtain an additional blood sample for an HIV test on the day participants pick up their study drug. A few participants test positive for HIV on that second test. At the time of the analysis, you remove those participants from the sample, in a modified intention to treat approach. That is perfectly appropriate.

Another example of a modified intention to treat is when you include in the analysis only those participants who received at least one dose of the study drug. Those who were randomized but never received any drug at all are excluded.

As Treated Analysis is used sometimes as a secondary look at the data when there have been many crossovers or low adherence to study interventions. It may be informative in regard to what happens when you actually receive the treatment. However, it is quite susceptible to bias, unless performed in a Non-Inferiority RCT. We discuss Non-Inferiority trials in a separate chapter.

6.13 Sequential Stopping Boundaries

All Randomized Clinical Trials should have an independent oversight committee, known as the Data and Safety Monitoring Board, or DSMB. The DSMB should be composed of independent academic researchers, and at least one experienced statistician. The DSMB examines periodically the event rates in both treatment arms, and any unanticipated problems that might arise.

The DSMB uses pre-specified stopping boundaries, also known as monitoring boundaries, and at each pre-specified look

they inform investigators whether they should proceed with the study or stop it. This protects study participants. Reasons to stop may be one of the following three:

1. **Proven Benefit.** The treatment is efficacious, and it is not ethical to prevent the control arm participants from receiving it.
2. **Probable Harm.** The treatment, unexpectedly, appears to be harmful and we must stop right away.
3. **Futility (no difference).** We have conducted the study for quite some time, recruited and followed many participants, and the event rates of the study arms still remain similar. It is extremely unlikely that we will be able to prove benefit.

There are many different ways to design Stopping Boundaries, but the underlying principle in all of them is that there needs to be an ethical decision in regard to when to stop the trial.

Whatever the method, most of them take into account that multiple comparisons between study arms require adjustment to avoid Type 1 Error increases.

In Fig. 6.3 we can see that the O'Brien-Fleming method compensates for multiple looks (avoids Type 1 error) by making it much harder to stop at the earlier interim comparisons, and then requires an ever-smaller difference between arms to stop at latter looks. The Haybittle-Peto method requires the same strength of evidence to stop at all interim analyses, and then requires less at the final analyses. The Pocock method always requires the same amount of information to stop at any time.

All those stopping boundaries may be classified as Symmetric, because the boundaries are mirror images of each other, and thus require the same amount of evidence to stop for benefit as for harm.

However, more recently we started also using Asymmetric Boundaries, which require less evidence to stop because of possible harm. When the outcome is severe it is better from an ethical perspective to use an Asymmetric Boundary because, as shown in the graphic below, it will detect potential harm to intervention participants faster than a Symmetric Boundary. In the Fig. 6.4 each red X mark depicts the Log-Rank statistic obtained at each of three interim analyses. On the left side of the graphic we see

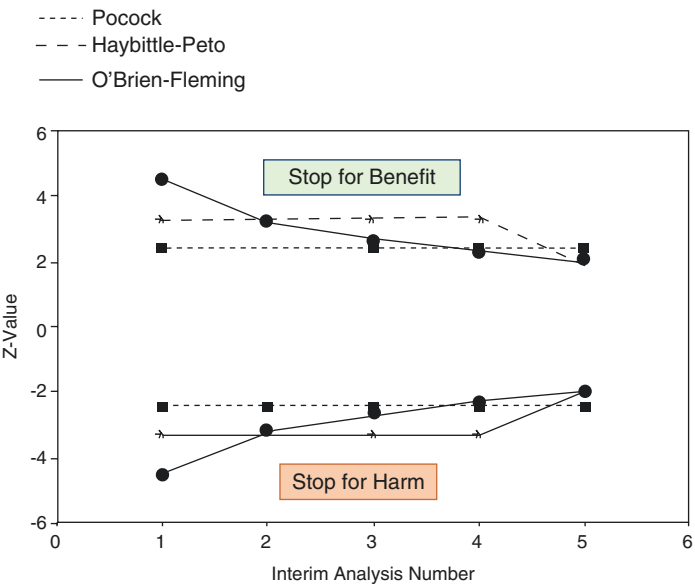


Fig. 6.3 Symmetric Stopping Boundaries. The O'Brien-Fleming method initially requires a larger difference between treatment arms to stop. The difference required becomes subsequently smaller

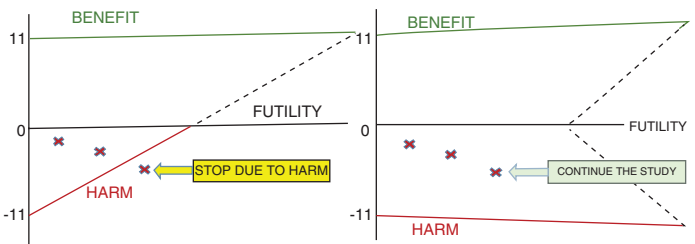


Fig. 6.4 Advantages of Asymmetric Boundaries. When analyzing the same data the asymmetric boundary (on the left) stops the study at the third interim analysis, thus reducing possible harm to participants

that an asymmetric boundary would be crossed at the third interim analysis, whereas a symmetric boundary (shown on the right) would allow the study to continue.

From an ethical perspective it seems clear that asymmetric boundaries should become the approach of choice.

A good example of how to use asymmetric boundaries was provided by the MADIT II study [Moss AJ et al. *N Engl J Med* 2002; 346:877–883].

MADIT II randomized patients with poor left ventricular function after a myocardial infarction to having an ICD implanted or continuing their usual management.

Very appropriately, no harm was expected (or tolerated) in the ICD arm. Accordingly, stopping for Benefit required far stronger evidence than stopping for Harm. The study was stopped early due to Benefit, and participants in the control arm were offered a defibrillator. Because of its shape, this type of sequential monitoring is known as a Triangular Design.

Figure 6.5 shows an example of a Triangular Design, similar to the one used in MADIT II. In this example the study was stopped because of observed benefit from the intervention at the third pre-specified analysis. You probably noticed that the Boundary for Benefit started at a value of 11 for the Log-Rank statistic, and that its slope is not entirely horizontal. Instead, it has a slight ascending slope that increases the value of the Log-Rank required to stop for benefit at each consecutive comparison. The reason for that slope is that we want to keep the probability of Type 1 Error at <0.05 as we perform repeated comparisons between the two study arms. We do that asking for a larger difference between arms each time we compare them.

You may have also appreciated another feature in Fig. 6.5: the serrated line that starts at 0 and goes up until crossing the boundary for benefit. That line represents the day-by-day estimate of the Log-Rank statistic, throughout the entire follow up. It was created after the study had already been stopped at the third analysis. It is quite informative because it shows that there was a consistent

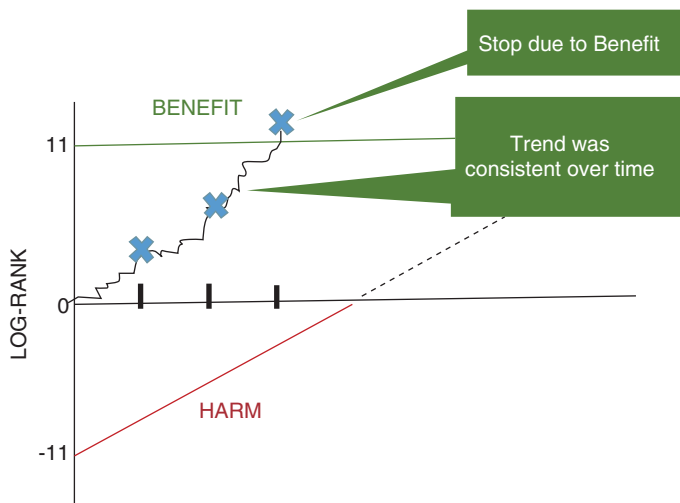


Fig. 6.5 Stopping Boundaries in MADIT II. The study was stopped at the third interim analysis because significant treatment benefit was ascertained

improvement in survival with the intervention. This consistency is very important because it confirms that the researchers did not arbitrarily stop the study when there was a convenient fluctuation in favor of the intervention. Given the concern that early stopping of clinical trials may lead to an overestimation of treatment benefit, this post factum day-to-day depiction of the Log-Rank statistic should be a mandatory feature for the publication of study results.

6.14 Improvements in Trial Monitoring

The main limitation of the trial monitoring using the log-rank or the Z statistic is that they do not include an assessment of the actual effect size or its confidence interval. In the future you may expect to see more studies that monitor the Relative Risk or the Hazard Ratio, and make plausible projections that help determine whether to continue or not. Figure 6.6 depicts a scenario in which

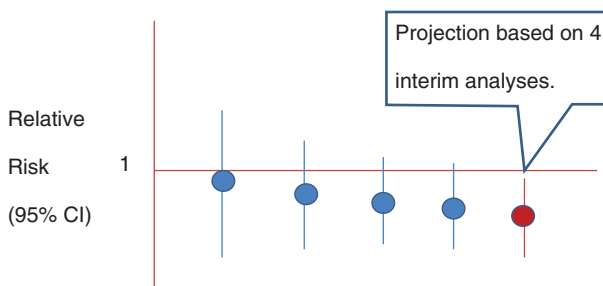


Fig. 6.6 Estimated Relative Risk. The Relative Risk measured at 4 interim analyses was used to estimate a projected value for the next step, which showed expectation of a statistically significant result

an informed decision to continue the study can be made, based on a projected Relative Risk and 95% confidence interval (shown in red) derived from data observed in four interim analyses.

6.15 Primary and Secondary Outcomes

As we already discussed, if we perform many comparisons between the treatment arms of a study we increase the probability of one of those comparisons yielding a false positive result. In order to limit the probability of Type 1 Error we emphasize the relevance of a single pre-defined Primary Outcome, or Endpoint, which we hope will be significant with a P -Value < 0.05 . In clinical trials jargon, we “spend all our alpha in the primary outcome assessment”.

For obvious reasons, the Primary Outcome should be clinically relevant, and our sample size should ensure adequate statistical power, to reduce the probability of Type 2 error as well.

Quite frequently, RCTs have a Composite Primary Outcome. For example, in the cardiovascular literature we frequently encounter the use of a Combined End Point that includes death, or non-fatal stroke or non-fatal myocardial infarction, whichever happens first. Other studies used a broader combination known as Major adverse cardiac events (MACE). MACE are frequently

defined as the composite of total death; MI; stroke, hospitalization because of HF; and revascularization, including percutaneous coronary intervention, and coronary artery bypass graft.

Composite outcomes are used because they increase the number of events we can analyze and thus increase statistical power.

However, several caveats apply, as we discuss next.

First, please remember to examine the results for each one of its components, to determine whether you have an improvement in all of them, or not. This is particularly important when softer events, such as revascularization, are combined with death.

Second, it should be remembered that composite outcomes count each participant who has multiple events only once (fatal or non-fatal, whatever event happens first). It is possible that two arms are very similar regarding a Composite Outcome because of frequent early rates of the softer non-fatal event, even if later on more participants in one arm went on to die. We could miss those deaths if they affect people who already had the softer event (i.e., had already been counted).

Third, whatever Primary Outcome the researchers have chosen, they must stick to it to avoid the suspicion that they went on a fishing expedition. You, however, have much greater freedom and may exercise discretion as an educated reader. For example, you should always examine the results regarding All-Cause Mortality, whether it was a Primary Outcome or not. You need to make sure that an effect on overall survival was not missed by the reported analysis.

If the risks of all possible bad outcomes (including death) are reduced by an intervention, that is quite reassuring because those risks are not competing with each other.

6.16 Hierarchical Testing of Secondary Outcomes

As we have discussed, we always choose one clinically relevant primary endpoint, because we must limit the probability of Type 1 Statistical Error. We reject the null hypothesis if we obtain a $P < 0.05$ for that outcome. However, we may have

important secondary endpoints. How can those secondary endpoints be analyzed without increasing the probability of Type 1 error?

Increasingly, researchers are using methods that allow them to test secondary endpoints without increasing the overall probability of Type 1 error. Several of those methods perform Hierarchical Significance Testing. Predetermining the order in which testing will proceed, through a clinical hierarchy, is a way to control the familywise (overall) Type 1 error. This is known as a Fixed Sequence Hierarchical testing approach. The basic rule is that we can only proceed to the next test if the null hypothesis is rejected for the current test. This preserves the overall probability of Type 1 Error at $<5\%$.

In its simplest form the primary outcome acts as a gatekeeper, and we only proceed to test secondary outcomes if the primary outcome was found to be significant, as shown in Fig. 6.7. We perform all the secondary analyses if the primary analysis yielded a significant result. The number of secondary endpoints should be very small, or a P -Value adjustment should be applied at the second step, to reduce the risk of Type 1 Error.

6.16.1 Stepwise Hierarchical Testing

We study a new drug for Congestive Heart Failure (CHF). We choose survival as our primary outcome, but we also want to determine whether this new drug can reduce CHF-related hospi-

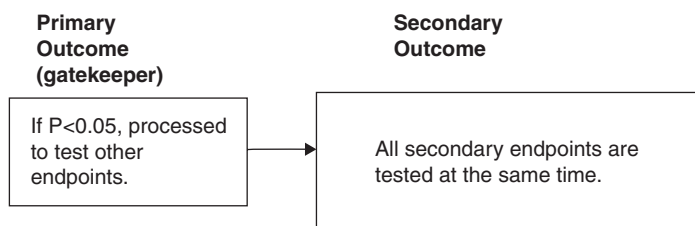


Fig. 6.7 Gatekeeping by Primary Outcome. We proceed to analyze secondary outcomes if the P -Value for the Primary Outcome is <0.05

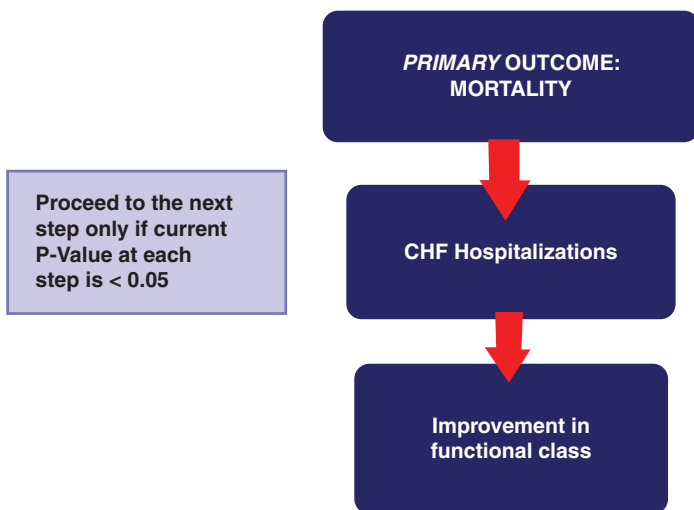


Fig. 6.8 Stepwise Hierarchical Testing. We proceed to the next step only if the current P -Value is < 0.05 . The initial step assesses our primary outcome

talizations, and improve symptoms as measured by functional class. We could perform a multi-step hierarchical testing approach, as depicted in Fig. 6.8. The gatekeeper system in this case stipulates that at each step we need to reject the null hypothesis in order to move to the next step. If we fail to reject the null hypothesis at any step, we must stop.

This type of hierarchical testing should be carefully designed to ensure that the initial steps are the most clinically relevant, and that adequate statistical power is provided by our sample size for each one of the steps.

6.17 Subgroup Analysis

Almost every paper reporting the results of an RCT contains, at the end, a figure (or table) that describes results of the Sub-Group Analysis. It is important to remember three things about Sub-Group Analysis, as follows

1. The subgroup analyses look for Interactions. As we discuss in the Basics of Biostatistics Module, an interaction (or effect modification) takes place when a third variable changes the relationship between the exposure and the outcome, without having to be associated to either of them. In the case of Sub-Group Analysis, a significant interaction means that a patient characteristic (such as being a diabetic) changes the way that patient responds to the treatment. In other words, if the treatment effect is significantly different in diabetics, as compared to non-diabetics, we have detected an interaction by diabetes mellitus. The *P*-Values in a subgroup analysis tables corresponds to the assessment of interactions across the categories of each characteristic.
2. A subgroup analysis is only exploratory. It generates hypothesis for research, but it does not confirm any hypothesis. In other words, finding an interaction tells us that we should carry future studies to confirm the presence of that interaction.
3. A subgroup analysis should be interpreted in the light of the overall findings and its own biological plausibility.

In particular, significant effects in isolated subgroups must be viewed with skepticism when the primary endpoint main analysis failed to reach statistical significance. In that case, Type 1 Error is likely, even if we make an adjustment for multiple comparisons.

On the contrary, once we have rejected the null hypothesis for the main analysis we would not be surprised to find out that the treatment is efficacious in all the subgroups we examine.

From a practical perspective, when you examine a Sub-Group Analysis figure please follow the four steps outlined below.

1. Look for Consistency with the Primary Comparison. Your initial frame of reference to examine the results is the point estimate for the Overall Comparison, our old friend the Primary Endpoint for the entire study. You should look up and down the graphic to see whether the point estimates for all subgroups are vertically aligned with the overall finding, or not.

As noted above, if the Overall Finding showed benefit from the intervention, you should expect most subgroups to show a trend towards benefit as well.

2. It does not matter whether the Confidence Interval (CI) for any subgroup crosses the Null Hypothesis line (e.g., the vertical line representing a Relative Risk of 1). Studies are not usually powered to look at effects by subgroups, and we expect to see wider CI's when we start dividing our sample into subgroups. Therefore, we can see some subgroup confidence limits cross the null hypothesis line, but that does not mean that those subgroups did not benefit from the treatment.
3. Compare effects between subgroups. If you identify any point estimate that seems to diverge from the Overall Comparison point estimate, then proceed to compare all the subgroups of that characteristic to each other (male versus female, smokers versus non-smokers, etc.). What matters is how the point estimate and CI's of each category/subgroup compares to the other categories. If the point estimate for any subgroup overlaps with the CI of the other category (or categories), we can be certain the *P*-Value for the interaction will be >0.05 . On the other hand, if the point estimates are far apart from each other and there is no overlap of the CI's, we can expect the *P*-Value for the interaction to be significant (<0.05).

Of note, a minimal overlap of the CI's may be associated with a *P*-Value <0.05 for the interaction. That happens because our methods to estimate Confidence Intervals tend to be conservative (they err on the side of giving us wider estimates, rather than narrower ones).

4. Determine whether it makes biological sense. In other words, any statistically significant finding must be supported by the Biological Plausibility of that finding. The ISIS2 investigators gave us a classic example of an implausible subgroup analysis that highlighted, with great sense of humor, the potential unreliability of multiple comparisons {The Lancet. 1988. 332:349-360}. They compared the benefit of Aspirin versus placebo after a Myocardial Infarction by astrological sign subgroups. Overall, Aspirin was significantly better than placebo. However, there was no statistically significant benefit (there was actually a trend towards harm by aspirin) if the patient was a Gemini or a Libra (Fig. 6.9).

GEMINI

The ruling planet Mercury is present in Sagittarius which imparts a dual influence for your health. if you suffer a heart attack you should not take any ASPIRIN.

Fig. 6.9 *Too many comparisons.* Or the unappreciated value of Astrology

6.18 Adaptive Clinical Trials

To understand why Adaptive Trials are considered a relatively novel approach in the twenty-first century we must look back to the twentieth century. Back then, substantial efforts were invested in ensuring trial transparency and preventing manipulation by investigators who were keen on having significant trial results, and might have biased their methods towards false positive findings.

The efforts to prevent that type of bias culminated in a system that required trial methods to be designed and described in their entirety, from randomization to analysis, in a public database such as clinicaltrials.org, before the first participant was recruited. Any departure from the a priori design had to be abundantly justified, and even with adequate justification it might be seen as a challenge to the study's validity. That system achieved its goal of maximizing transparency and minimizing false positive findings, but it was limited by its inflexibility and inability to adapt to rapidly evolving clinical scenarios. Adaptive trials maintain the transparency provided by a priori descriptions, but also allow us to make certain changes in our methods, in a “learn as we go” approach. Adaptive strategies may be implemented at all stages of a trial from randomization to analyses, as shown in Fig. 6.10. The first step in this flowchart is Participant Enrollment. The yellow shaded boxes show some of the optional adaptive steps that we may take as the study proceeds.

An outstanding example among adaptive trials is the Randomised Evaluation of COVID-19 thERapY (RECOVERY) trial, launched in March of 2020 through a collaboration between

until enough follow-up data from the control group was available. As we discuss in the section Assessing a Negative Trial, one of the essential items when we calculate sample size requirements is to predict the incidence of the outcome in the control (Standard Treatment) arm. If we overestimate that incidence our calculations will result in an inadequately small sample and insufficient statistical power. Of note, the RECOVERY researchers' flexible method, which ensured that their sample size estimates were based on reliable data, was not the predominant approach at the time. It was one of the many innovative dynamic design features that made RECOVERY a resounding success.

6.19 Checklist for a Randomized Clinical Trial

- Was the study registered in advance in clinicaltrials.org or equivalent registry?
- Did the study comply with the CONSORT guidelines?
- Is the trial's research question the most relevant in this area?
- Who were the study participants? Are they similar enough to my patients?
- Was randomization adequate?
- Was the intervention double-blinded? Was there involuntary un-blinding?
- How many participants were lost to follow-up? How many participants crossed over?
- Were events ascertained in a complete and blinded manner?
- Was the analysis performed on an intention to treat basis? Would modified intention to treat, or "as treated" analyses be informative?
- What was the treatment effect estimate? What was the number needed to treat? Was all-cause mortality reduced?
- What were the adverse effects? What was the number needed to harm?
- Was there any evidence of effect modification in sub-group analysis? If so, was it biologically plausible?

- Is there information regarding cost, or cost-effectiveness?
- If there were prior studies, how do these results compare to previous results? Is a meta-analysis feasible?

6.20 Assessing a Negative Clinical Trial

The first question we must answer when assessing a study that failed to reject the Null Hypothesis is whether it had adequate statistical power. We need to determine whether their sample size estimate assumptions were accurate. Yes, you do need to overcome your phobia and read the Statistical Methods section of the paper to see whether it all worked as expected, or not.

Let us examine the ingredients of the sample size recipe.

- Type 1 Statistical Error. Most studies have one major outcome, for which they wish to reject the Null Hypothesis with a $P\text{-Value} < 0.05$.
- Type 2 Statistical Error. The convention in this case is that the power should be at least 80%. In other words, Type 2 Error should be 20% or less.
- Outcome incidence in the Control Arm. It is easier to show a reduction of something that happens quite often in the Control Arm. Similarly, if the event rate in the Control Arm is low we will need a larger sample size to detect the same amount of reduction.
- Expected benefit. We usually call this the Effect Size Estimate. If we are too optimistic about the therapeutic benefit from our treatment our sample size will be too small, and our power too low.
- Crossovers. All crossovers reduce our statistical power and result in the need for a larger sample size. Higher crossover rates are particularly predictable when: (1) the evaluated treatment is available outside of the trial; (2) a double-blinded design is not feasible; and (3) the treatment is perceived as highly efficacious.
- Losses to follow-up. They would also reduce our power and should be compensated for through a larger sample size.

- Intervention fidelity. If the participants do not receive the full benefit of the intervention our power will be reduced. For example, if we are assessing whether a novel percutaneous intracoronary stent implantation reduces negative cardiac outcomes we should ensure that the interventional cardiologists in our trial are fully trained in the use of the new stent.

6.21 Checklist for a Negative Clinical Trial

- Did the researchers recruit the number of patients they intended to?
- Was the event rate in the control group as high as expected?
- Were losses to follow up as infrequent as expected?
- Were crossovers as infrequent as expected?
- Was the intervention carried out as expected (intervention fidelity)?

→ If the answer is “NO” to any of the above questions, the possibility of lower-than-expected power should be considered. But remember: if the treatment were efficacious, we expect to see at least a trend towards improvement in the intervention arm.

In addition to assessing the statistical power you should consider the following items to determine whether useful information may still be gained from a negative trial, and whether further studies may be warranted:

- Was the intervention an optimal representation of this type of treatment?
- Was the primary outcome appropriate?
- Was the population appropriately identified and recruited?
- Was the trial appropriately conducted?
- Should a non-inferiority study be considered?
- Is there a biologically plausible finding in subgroup analyses that merits further investigation?
- Were the data optimally analyzed?
- Is this study of good enough quality to be included in a meta-analysis?

Non-inferiority Clinical Trials

7

7.1 Why Do We Need Non-Inferiority Trials?

Placebo-controlled studies are no longer ethical when there is already a treatment for the pertinent disease process. We perform a non-inferiority trial when we expect the new treatment to have similar efficacy as compared to available treatment, and to have an advantage in at least one of the following: safety, convenience, tolerability, acceptability, and/or cost.

Obviously, if we have preliminary evidence that the new treatment has greater efficacy than what is already available, a traditional superiority design should be used.

7.2 A Quick Reminder about Superiority Trials

Superiority clinical trials are the traditional study design. We are all familiar with them. In fact, most lay people assume that all randomized clinical trials are done to test superiority. Let us do a

quick review so we can later make a contrast with non-inferiority designs.

A superiority trial assesses whether the new treatment is better than the standard of care. Thus, the study is conducted to show that the new treatment is better in reducing the risk of a bad outcome than the available treatment (or better than placebo, if there is no treatment yet). With the appropriate design, which includes participant safety monitoring and preplanned interim analyses, superiority trials seldom are ethically controversial.

7.3 Hypotheses in Superiority Trials

Let us consider a disease that causes substantial mortality. Naturally, we would like to identify new treatments that reduce the risk of death. When we assess such a new treatment in a superiority trial we hope that the mortality will be lower in the Treatment Arm than in the Control Arm. However, as we discussed in the Basics of Biostatistics module, we do many counter-intuitive things in frequentist statistics. For example, our Study Hypothesis, which we call H_1 , includes the possibility that mortality is increased by our new treatment, or Intervention. Not something we ever desire, but something that we must consider.

H_1 = Mortality in the Intervention Arm is less than, or greater than the mortality in the Control Arm.

You should also remember that, oddly enough, our statistical tests do not attempt to confirm our hypothesis. Rather, they aim at rejecting the Null Hypothesis (H_0) of no difference in mortality between study arms.

H_0 = Mortality in the Intervention Arm is not different from the mortality in the Control Arm.

When we analyze our data we perform a Null Hypothesis Significance Test, and the main output is the *P*-Value.

The *P*-Value represents the probability of observing a difference in mortality as large as the one we detected, or larger, if the Null Hypothesis were true. If the *P*-Value is small enough (usually <0.05), we reject the null hypothesis because it's unlikely to be true given our data. Other possible but much less satisfactory outcomes include: (a) a trend in mortality reduction or increase that did not reach statistical significance; (b) a statistically significant increase in mortality.

7.4 Type 2 Error in Superiority Trials

As we have previously discussed the following occurrences can make the outcome rates more similar between the study arms, thereby lowering our statistical power. In other words, they all increase the probability of Type 2 Error (i.e., there is a greater probability of a false negative study).

- Excessive losses to follow-up in both arms
- Too many crossovers (because regardless of participants crossing over we must analyze them in the group they got randomized to: Intention to Treat)
- Lower than expected event rate in the control arm

7.5 A Whole New World: Hypotheses in Non-Inferiority Trials

A new treatment is considered as non-inferior to the available, older, treatment if it is better, equal, or “not significantly worse” than the standard of care in regards to the main outcome of interest. For example, we c.

We accept the possibility of some efficacy loss as being “not significantly worse”, and how much of a loss we accept as part of non-inferiority is a crucial ethical and clinical issue.

Using effects on mortality as an example, our non-inferiority hypothesis is that mortality in the Treatment Arm is not significantly larger than in the Control Arm. Therefore, our Hypothesis (H1) now encompasses a reduction, but also a non-significant increase in mortality.

H1 = Mortality in the Intervention Arm is less than, or similar to the mortality in the Control Arm.

In turn, the Null Hypothesis (H0) we aim to reject is that mortality is increased by the new Treatment by more than a pre-defined value, which we call the non-inferiority margin.

H0 = Mortality in the Intervention Arm > Mortality in the Control Arm.

The Null Hypothesis has changed, as compared to superiority. We thus need a new approach to significance testing. Let us start with the differences in regards to the confidence interval for our test statistics. The test statistic may be the same as for superiority analysis (e.g., Relative Risk), but in regards to the confidence interval we care only about the upper confidence limit, which is the side of the Confidence Interval pointing towards an increase in mortality caused by the Treatment Arm.

If the upper confidence limit of our test statistic crosses the non-inferiority margin, we fail to reject the null hypothesis.

Two different test statistics have been used in non-inferiority trials: the Relative Risk (RR) and the Absolute Risk Difference

(ARD). It should be noted that the ARD has recently been favored by researchers, and many studies have used that method. This is consequential because absolute risk changes are not always easily understood, and thus the clinical implications of “statistically significant” results may remain rather obscure to many clinicians.

Whatever the method, RR or ARD, it is essential to understand how much efficacy loss any given margin accepts as part of the statistical definition of non-inferiority. Non-inferiority margins may be arbitrarily defined, and a “statistically significant” finding of non-inferiority may accept a loss in efficacy that is, actually, clinically unacceptable. In general, a wide non-inferiority margin should be viewed as unethical, because it may deem “non-inferior” a new treatment that is actually less efficacious than the older one.

7.6 The Relative Risk (RR) Non-Inferiority Margin

As usual, the RR is estimated dividing the treatment event rate over the control event rate.

$$\text{RR} = \text{Event Rate in Treatment Arm} / \text{Event Rate in Control Arm}$$

If we set the non-inferiority margin at a $\text{RR} = 1.10$, that means that to claim non-inferiority the upper confidence limit of the RR cannot be greater than 1.09. That pretty much guarantees that the RR point estimate will be at, or below 1. This is really compatible with non-inferiority, and thus it is ideal from an ethical perspective.

In general, I would recommend considering $\text{RR} = 1.25$ as the highest margin we can accept while clearly safeguarding the clinical meaning of non-inferiority. As we discuss later, one of the risks of testing the ARD is that the chosen margin may translate into unacceptable lenient margins in RR units.

In Fig. 7.1 we can see, from top to bottom, the three possible outcomes for a non-inferiority study:

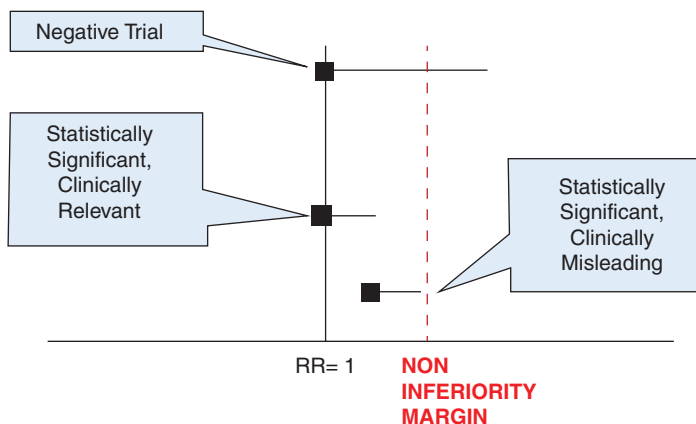


Fig. 7.1 Possible outcomes in non-inferiority. From top to bottom the graphic shows the results of a negative study; then a statistically significant and clinically relevant study; and finally, a statistically significant but clinically misleading study

- Top: The upper confidence limit of the test statistic crosses the non-inferiority margin and the study fails to reject H_0 . This is a negative trial.
- Middle: The point estimate is compatible with very similar efficacy in both arms, and the upper confidence limit does not cross the non-inferiority margin. This result is both statistically significant and clinically relevant.
- Bottom: The point estimate suggests an increase in mortality in the Treatment Arm, but the lenient non-inferiority margin (far away from the RR of 1) allows the upper confidence interval to remain within the “non-inferiority” region. The study researchers (and the drug/device manufacturer) may claim success, but this is not a great outcome for patients or for society, as we discuss later in this module.

7.7 The Absolute Risk Difference (ARD) Non-Inferiority Margin

The ARD has become the method of choice to assess non-inferiority in many studies. It is calculated subtracting the control event rate from treatment event rate. This measures the absolute change in risk introduced by the treatment.

$$\text{ARD} = \text{Event Rate in Treatment Arm} - \text{Event Rate in Control Arm}$$

A decimal point notation is used for both event rates and the ARD is thus usually expressed in percentage point units. Let us consider what a percentage point actually means. They are different from percent changes. They sound almost the same, but they are not.

For example, consider a study in which the mortality rate in the control arm is 40%

- A 10 percentage point increase in mortality would give us a 50% event rate in the treatment arm.
- In turn, a 10% increase will give us a 44% event rate in the treatment arm.

An ARD of zero means the event rates are identical—same risk of the outcome in both study arms. That corresponds to a Relative Risk of 1.

A positive ARD >0 indicates higher risk of the outcome in the treatment arm, and we call the difference an Absolute Risk Increase (ARI). In turn, if the ARD is a negative number the risk of the outcome is lower in treatment arm, i.e. the treatment is beneficial. We call that difference an Absolute Risk Reduction.

We use the ARI to estimate the Number Needed to Harm (NNH), which is the number of patients who need to be switched

to the new treatment to observe one additional poor outcome in our sample, over a period of time similar to our study's duration.

$$\text{Number Needed to Harm} = 1 / \text{Absolute Risk Increase}$$

7.8 Both Relative and Absolute Risk Are Clinically Important

No matter what method the researchers use to assess non-inferiority, the clinical meaning of the non-inferiority margin and the final study results should always be assessed by us examining the Relative Risk, the Absolute Risk Difference, and the Number Needed to Harm.

Of note, effects on the absolute risk are very important from a societal perspective, but our patients are more likely to understand (and be interested in) relative risk changes.

7.9 Type 1 Error in Non-Inferiority Trials

It is extremely important to remember that what used to be Null Hypothesis in superiority trials (same event rates in both study arms) is now part of the Study Hypothesis.

All the factors and biases that increase the probability of Type 2 Error in superiority studies increase the probability of Type 1 Error in non-inferiority. They tend to decrease the difference in event rates between the study arms over follow-up, making them appear more similar in regards to their outcomes.

In addition, the choice of a Control Arm has to mimic previous landmark trials and reflect the optimal standard of care in regards

to patient population, doses, duration of treatment, etc. If the Control Arm receives suboptimal care a non-inferiority finding would be misleading.

Furthermore, the outcome incidence in the Control Arm should be similar to that in previous studies and reflective of risk in our population. Lower than expected event rates may reflect a lower risk in our study sample, leading to treatment effects that may not be representative of what will happen in our population. That is another possible cause of False Positive non-inferiority results because the Treatment did well simply because we recruited healthier patients.

A false positive finding of non-inferiority may occur because:

There are too many crossovers and/or losses to follow up.

The Control Arm does not reflect the best standard of care.

The study sample has a lower risk of the outcome than our population of interest.

7.10 As Treated and Intention to Treat Analysis

In non-inferiority trials we should always perform both As Treated (AT) and Intention to Treat (ITT) Analysis. This is another major difference from superiority trials, which are required to perform an ITT Analysis as their primary approach.

We hope that the number of participants who cross over will be small enough to make both As Treated and ITT results very similar.

7.11 Superiority Analysis in a Non-Inferiority Trial

It is perfectly appropriate to plan for a superiority analysis in a non-inferiority trial, to be performed only if null hypothesis for non-inferiority has been rejected. In other words, it is fine to examine whether the Treatment is superior to the Control only after finding that it is non-inferior. Remember: non-inferiority includes having the same efficacy and having better efficacy than the Control arm.

However, some researchers erroneously proceed with a superiority analysis even though they failed to reject the Null for non-inferiority. That is nonsensical and potentially misleading, as we will discuss in one of our case examples, next.

7.12 ABSORB III: A Non-Inferiority Study that Used an Absolute Risk Difference Margin

The ABSORB III study compared an everolimus-eluting biore-sorbable vascular scaffold (Absorb) to an everolimus-eluting cobalt-chromium stent (Xience 41) in patients with angina pectoris. *Everolimus-Eluting Bioresorbable Scaffolds for Coronary Artery Disease*; Ellis SG et al.; *N Engl J Med* 2015; 373:1905–1915. Everolimus stents were already established as efficacious. The primary end point was target-lesion failure at 1 year, and an absolute risk difference (ARD) method was chosen to assess non-inferiority.

Event rate in the Xience arm was expected to be 7%. A non-inferiority margin of 4.5 percentage points (0.045 in absolute notation) was chosen for the absolute difference in risk. Let us examine what was a priori viewed as statistically non-inferior. Their margin meant that an upper confidence limit for the ARD up to 0.04 would have been “non-inferior”. That was compatible with an upper confidence limit for the Number Needed to Harm (NNH) of only 25:

$$\text{NNH} = 1 / 0.04 = 25$$

In Relative Risk (RR) units the non-inferiority margin was 1.64:

$$RR = (0.07 + 0.045) / 0.07 = 1.64$$

That ARD threshold meant that a result with a confidence limit of the NNH of 26, and/or an upper confidence limit for the relative risk of 1.63 was “acceptable”. Clearly, this was a very permissive non-inferiority margin, in terms of what it accepted in possible numbers needed to harm, as well as in possible relative risk increments.

The clinical implications of a non-inferiority margin should always be assessed, and the results evaluated as Relative Risk and Number Needed to Harm.

Let us now examine the ABSORB III study results. Target-lesion failure at 1 year happened in 102 of 1313 (7.8%) participants in the Absorb arm, and in 41 of 677 (6.1%) participants in the Xience arm. The absolute difference (95% confidence interval [CI]) was 1.7 (0.5 to 3.9) percentage points, $P = 0.007$ for non-inferiority.

The authors concluded, quite correctly, that the findings were “within the pre-specified margin of non-inferiority.” That assertion aside, do you agree that the Absorb device is similar enough in efficacy to the Xience stent? What do their results mean in terms of RR and NNH?

The RR (95% CI) was 1.28 (0.90 to 1.82). The RR point estimate suggests that there was an estimated 28% greater risk of target lesion failure with Absorb. I doubt my patients would view that as evidence of non-inferiority.

From a societal perspective, the NNH (95% CI) was 59 (25 to –136). Considering that about half a million Americans have been estimated to undergo a percutaneous coronary revascularization every year, a generalized switch to the Absorb device might cause thousands of additional target-lesion failures every year. *Heart Disease and Stroke Statistics—2016 Update. A Report from the American Heart Association. Circulation 2016;133:e38–e360.*

ABSORB III is just one of many studies that have applied a lenient, permissive ARD non-inferiority margin. However, its publication in a prestigious medical journal represented, in my view, a serious failure of the peer review and editorial oversight processes. Our collective fascination with new technologies should not obstruct a careful review of trial methods and results.

7.13 Is a Non-Inferiority Trial Acceptable when the Outcome Is Death?

OPTIMAAL, a relatively old non-inferiority, trial still provides an excellent example of the ethical issues raised when we perform a non-inferiority trial to assess a new treatment when the existing treatment is already known to reduce mortality risks. OPTIMAAL compared losartan, then a new drug, to captopril, a drug already known to reduce mortality risk in patients with heart failure after a myocardial infarction.

The researchers defined the non-inferiority margin as a relative risk of 1.1. This margin was appropriately very narrow, and meant to ensure that a non-inferiority margin would be obtained only if losartan really provided very similar therapeutic benefits, as compared to captopril. However, was that enough to protect the welfare of their trial participants?

The study had several features that were greatly problematic, as follows:

1. Study participants had to tolerate captopril to participate, meaning that participation implied risking the loss of proven benefit from captopril if randomized to the losartan arm.
2. The ongoing monitoring used stopping rules that were created to ensure enough power to detect non-inferiority, and potentially tolerated additional participant deaths as the study progressed. There may have been fifty additional deaths in the losartan arm, as compared to the captopril arm, by the end of the study. See *Who protects participants in non-inferiority trials when the outcome is death?* Palmas W. *Research Ethics* 2018, Vol. 14(1) 10–15 It seems obvious that this design feature violated the Beneficence principle.

3. It is unclear whether the consent process respected the Autonomy principle, because the consent forms have not been published and we cannot ascertain whether the possibly higher risk of death was clearly explained.
4. In spite of the failure to reject the null hypothesis for non-inferiority the authors proceeded to perform a superiority analysis, and provided the superiority analysis results in the abstract of their publication. This error was even more egregious given that the abstract is the most impactful part of any manuscript.

In summary, OPTIMAAL should be remembered as a sobering example of the dangers of performing a non-inferiority trial when the outcome is death.

7.14 Checklist for Non-Inferiority Trials

- What was the non-inferiority margin? This is the threshold for non-inferiority significance (null hypothesis line)
- How much potential efficacy loss did the non-inferiority margin accept? Was it clinically and ethically appropriate? Always examine both relative and absolute risk effects.
- If non-inferiority was examined solely through the Absolute Risk Difference, what were the results in terms of Number Needed to Harm, and Relative Risk (and their 95% confidence intervals)?
- Was the quality of the control arm a good reflection of the best available treatment (standard of care)?
- Were there losses to follow up, crossovers, and/or low intervention fidelity? They all increase the probability of type 1 error (as opposed to superiority RCTs, in which they increase the prob. of type 2 error).
- Was the event rate in the control arm as high as expected, as described in the Statistical Methods section? If it was too low that increases the risk of Type 1 error.
- Was an “as treated” analysis provided, along the usual “intention to treat” analysis?
- What is the clinical significance of the results? Do they apply to my patient?

Bayesian Analysis of Clinical Trials

8

8.1 Limitations of Conventional Statistics

The dominant approach to biostatistics is still the one widely adopted since the beginning of the twentieth century. At its core is the Null Hypothesis Significance Testing, which generates a P -Value. Let us briefly review again what that means, through an example.

An international group of investigators performs a clinical trial, called EOLIA, to assess the efficacy of venovenous Extracorporeal Membrane Oxygenation (ECMO) in patients with severe Acute Respiratory Distress Syndrome (ARDS), as compared to conventional treatment. Their hypothesis is that ECMO will reduce mortality at 60 days. To analyze with conventional statistics their results they must first define the Null Hypothesis, which states that there is no mortality difference at 60 days between ECMO and the control group. After that, they may proceed with the trial to see whether the results allow them to reject the Null Hypothesis.

When they complete their study the EOLIA researchers find a mortality rate of 35% in the ECMO group and 46% in the control group. That corresponds to a relative risk of 0.76; 95% confidence interval [CI] 0.55 to 1.04. The P -Value for the mortality difference is 0.09.

What does that mean? It means that the estimated probability of seeing a difference in mortality as large as the one observed, or larger, if the null hypothesis were true is 9%. A simplistic interpretation of the results would just state: “The P-Value is not significant and therefore the ECMO group did not do significantly better than the control group”. But you do not care for simplism and are impressed by the relative risk of 0.76, which corresponds to a relative risk reduction of 24%, and by the fact that the upper confidence limit of 1.04 is barely beyond 1. Your inquisitive nature makes you wonder, “What is the probability that ECMO reduces mortality, based on the results of the EOLIA trial?”

Can the Confidence Interval help? A 95% Confidence Interval means that if you were to perform similar trials numerous times, you expect that the confidence boundaries will contain the true population value 95% of the times. That is a very interesting concept, but a likely finding in multiple theoretical repeats is a rather Platonic concept, and does nothing to help answer your question. Maybe you should consider a Bayesian approach and obtain probabilities based on the EOLIA trial.

8.2 Similarities with the Diagnostic Process

Clinicians use Bayesian methods in their everyday work to diagnose disease. Before performing a diagnostic test, they assess the pre-test probability, or prior probability, of the disease. After performing the diagnostic test, they integrate the test results with the prior probability, and thus they obtain a posterior probability (post-test probability) of the disease being present.

Similarly, we can interpret the results of a clinical trial, given the prior probability and the study results. Bayesian estimates can provide answers to questions such as: “What is the probability that the risk of the outcome is reduced by the new treatment?”, and “What is the probability that the risk of the outcome is reduced by 10% or more by the new treatment?” If you are a pessimist, you may wonder “What is the probability that the new treatment increases mortality?”

8.3 Prior Probability, Data, and Posterior Probability

We start by determining what, if any, information we have that could inform the Prior Probability estimate that the new treatment is, indeed, better than the standard of care.

When we carry out the clinical trial, its results will be our Data.

The integration of the Prior Probability with the Data through Bayesian methods gives us the Posterior Probability of the new treatment being better than the standard of care.

The Posterior Probability has a point estimate for the Relative Risk or the Hazard Ratio, and a 95% Credible Interval, the Bayesian equivalent of a Confidence Interval. In addition, it has an underlying distribution that allows us to determine the probability of any effect of interest, such as Relative Risk Reduction of 10% or more.

8.4 Types of Prior Probability

Prior Probabilities can be classified into three big categories, based on our expectation of seeing a positive trial result:

Uninformative Prior: We do not have any prior studies to inform us, and we thus will take the current trial result (our Data) as the source of information to determine the Posterior Probability. Please keep in mind that the Uninformative Prior is also known in the literature as Non-Informative, Minimally Informative, and Vague.

Optimistic Prior: In this situation we have previous information, such as older studies, that create a certain expectation of having a positive result.

Pessimistic Prior: The available knowledge makes it unlikely that we will have a positive trial result. Of note, a clinical trial should not be carried out if a pessimistic prior is a realistic consideration—it would be clearly unethical. This makes pessimistic priors nonsensical for the analysis of clinical trials.

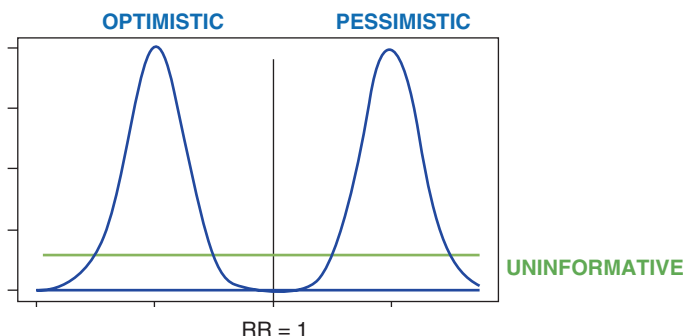


Fig. 8.1 Types of prior probabilities. The green flat line depicts the uniform distribution corresponding to the uninformative prior

Figure 8.1 depicts the three types of prior probabilities, using the relative risk as an example. The uninformative prior, in green, has a flat distribution—it assumes no prior knowledge, provides a naïve approach to the data, and will provide posterior probability results very similar to those of conventional frequentist statistics. The optimistic prior has a probability density that peaks at a point consistent with a significant risk reduction, whereas the opposite is true for the pessimistic prior.

8.5 Prior Probabilities and Clinical Common Sense

The Uninformative Prior should be, in most cases, the method of choice to analyze clinical trials because it is consistent with the principle of equipoise, it renders a Posterior Probability that is based mostly on the current trial results and gives us Credible Intervals that are extremely similar (often identical) to frequentist Confidence Intervals, thus avoiding controversy with old-fashioned statisticians. You may wonder whether this approach is appropriate when prior studies have already tested the same question. The answer is a resounding Yes, because you may still perform a subsequent meta-analysis to formally integrate all available information. In fact, a meta-analysis would be appropriate because

it would examine the validity of the evidence from each trial before meta-analyzing their data in a weighted manner.

The routine use of Optimistic Priors should be discouraged unless there are very well designed previous studies that have clearly showed efficacy of the treatment at hand. Optimistic Priors that are not based on a conservative assessment of good-quality historical results can open the door to criticism of Bayesian methods, because they may render Posterior Probabilities that show a benefit not suggested by the most recent Data at hand (as we discuss in an example below).

What about Pessimistic Priors? Some statisticians believe that including a Pessimistic Prior as one of their analytic models is fair, more exhaustive, and informative. I disagree. Prior probabilities should reflect at least a modicum of true prior belief. It would be grossly unethical to conduct a trial if there is a strong expectation that the new treatment will be worse than the standard of care. In addition, the Uninformative Prior will allow us to determine, without any bias, whether the new treatment is any worse than the standard of care, based on the study results. For those reasons, the Pessimistic Prior is nonsensical in the clinical trial setting.

8.6 The Use of Excessively Optimistic Priors Should be Avoided

Let us consider the case of a clinical trial that showed identical outcomes in both study groups, giving us a Relative Risk (RR) of 1. We can think of our RR result as having an underlying probability that peaks at a $RR = 1$, as depicted in Fig. 8.2.

Figure 8.3 shows how the posterior probability estimate can be excessively influenced by an extremely optimistic prior. The prior is represented by the blue distribution, whereas our result, the data, is still represented by the red distribution.

We went from the data showing no benefit to the distribution depicted in green—that of a posterior point estimate of $RR = 0.8$. The optimistic prior “pulled” the posterior towards an inflated perception of treatment benefit.

This example is intentionally exaggerated. It is meant to exemplify why excessively optimistic priors should be avoided.

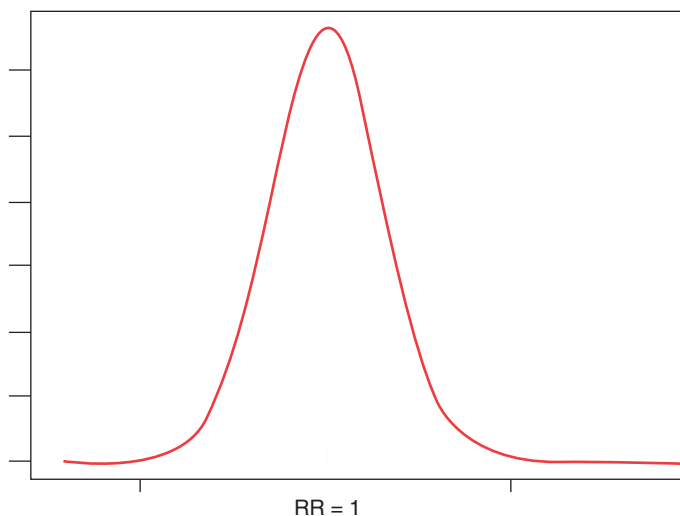


Fig. 8.2 Distribution for a Relative Risk = 1. The distribution peaks at our effect estimate

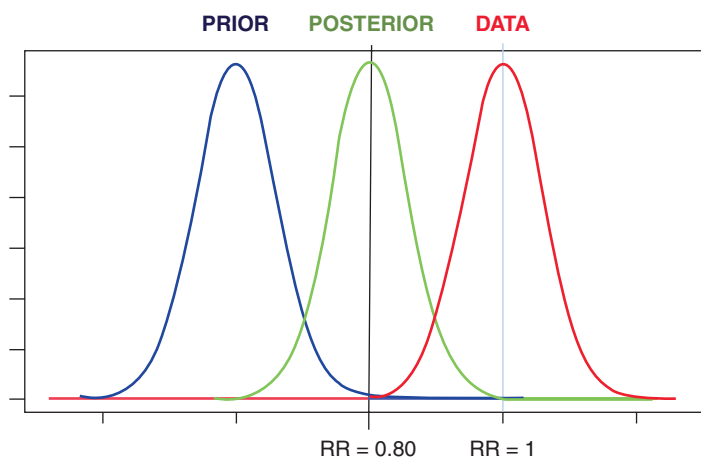


Fig. 8.3 Effect of an optimistic prior. The resulting posterior suggests significant efficacy, which was not suggested by the actual data

8.7 Advantages of the Uninformative Prior

The uninformative Prior, in turn, gives as a Posterior Probability based only on the Data, as shown in Fig. 8.4.

The Credible Intervals of the RR will be almost identical to frequentist Confidence Intervals, thus pre-empting any controversy with old-school frequentists. This shows why the uninformative prior is a better choice than an optimistic one when analyzing clinical trials.

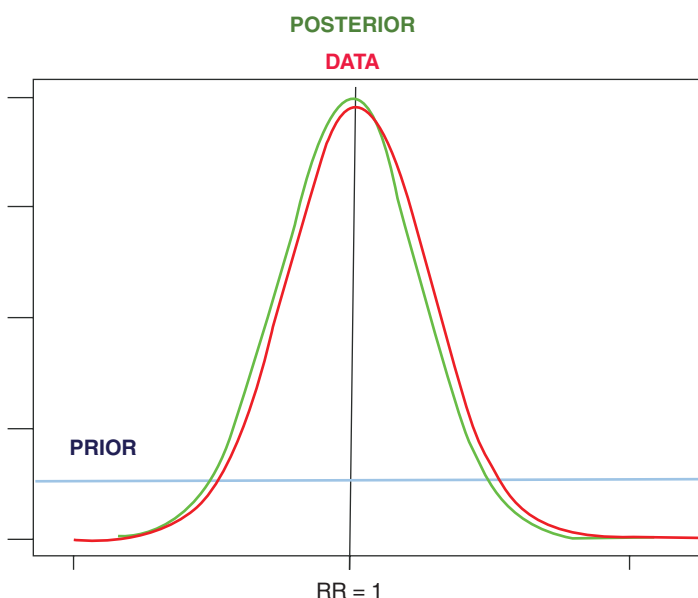


Fig. 8.4 Effect of an uninformative prior. The resulting posterior is determined completely by the data

8.8 Navigating Statistical Lingo in the Methods Section

When you read the methods section of a study that used a Bayesian analysis the unfamiliar technical language may be intimidating, particularly because it is usually redacted for brevity, rather than readability. There are a couple of things you should look for in there, as follows:

1. As we have extensively discussed, we would be happier if the authors used an Uninformative Prior. If they ran extensive modelling that included Optimistic, Uninformative, and Pessimistic Priors, **it is our prerogative as readers to pay particular attention to the Uninformative.**
2. For Bayesian modelling to work the Prior needs to follow a distribution that is similar to that of the result. We describe that congruency as being **Conjugate.**
3. The statistical process that integrated the Prior with the Data is usually described in the same sentence, very commonly as a simulation that sampled randomly from a given Probability distribution, such as a Markov Chain Montecarlo method. They work through repetitive random sampling until the estimates given by the sampling become acceptably similar to each other, reaching what we call **Convergence.**

8.9 Clinical Interpretation of Posterior Probabilities

Let us re-examine the results of the EOLIA trial. It reported a mortality rate of 35% in the ECMO group and 46% in the control group. That corresponded to a RR of 0.76; 95% CI of 0.55 to 1.04, $P = 0.09$.

If we model the posterior probability of the RR using an uninformative prior, using the method described by Wijeyesundera and

colleagues [Wijeyesundera, D. et al. “Bayesian statistical inference enhances the interpretation of contemporary randomized controlled trials.” *Journal of clinical epidemiology* 62 1 (2009): 13–21] we obtain the following:

$$RR(95\% \text{Credible Interval}) = 0.76(0.56 \text{ to } 1.04)$$

These Bayesian statistics are almost identical to the frequentist results, but they add estimated probabilities of treatment effects. The normalized probability distribution underlying those values is shown in Fig. 8.5 (the graphic uses log-transformed values to approximate a normal distribution).

The probability that the RR is 0.90 or lower, depicted in red in the graphic, is estimated as 86%. In other words, the probability that the ECMO group has a 10% reduction, or greater, in the mortality risk is 86%. The probability that ECMO reduces mortality by any amount is 96%.

How do you think a patient would react to those probability estimates? Would they care about the *P*-Value of 0.09?

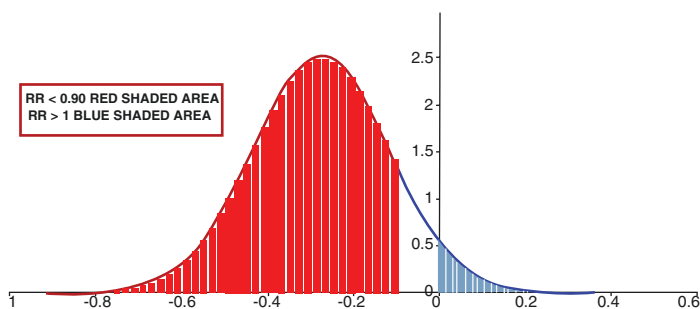


Fig. 8.5 Posterior probability of Relative Risk (RR). The probability of $RR = 0.90$, in red, was 86%, whereas the blue-shaded area depicts the probability of a $RR > 1$, which was 4%

8.10 Irrelevance of the Null Hypothesis

Surprisingly, some people examine Bayesian models and proceed to propose an interpretation of the results based on Null Hypothesis rejection areas, just like in the older frequentist approach. This is completely unnecessary because the Bayesian posterior probabilities are informative by themselves and deal with the Hypothesis directly, without any need to consider the Null Hypothesis. Please, do not look for a combustion engine under the hood of your electric car.

8.11 The Added Value of Bayesian Methods

As noted above, if we use an Uninformative Prior to analyze a large clinical trial our Bayesian point estimates (and their credible intervals) will closely resemble frequentists results. This should pre-empt criticism of Bayesian methods as arbitrary and subjective. Why bother, then, if the output will look very much like the frequentist result? Because the Bayesian methods add probabilities that are intuitively understood by clinicians and maybe shared with patients.

With current technology and openly shared software the computational issues are no longer a barrier to Bayesian methods. Bayesian analyses should be prospectively considered for all clinical trials, in conjunction with conventional frequentist analyses.



9.1 Basic Definitions

Health economics has the overarching goal of maximizing population health through the most efficient use of resources. Of note, health is understood in the broad sense of its meaning, as defined by the World Health Organization.

World Health Organization: “Health is a state of complete physical, mental, and social well-being, and not merely the absence of disease or infirmity.”

The first essential step in a cost-effectiveness evaluation is to define the key clinical question and the appropriate scope of the evaluation (www.nice.org.uk). An evaluation would be desirable if:

- There is uncertainty or disagreement about best practice.
- There is potential to improve important health outcomes and/or make better use of health resources.
- There is potential to reduce health inequalities and systemic bias.
- The evaluation is likely to improve clinical practice.

Once the relevance and scope of the question have been ascertained, the following items should be defined:

Perspective We usually favor a societal perspective—one that maximizes health gains for the largest number of people. All costs should be included regardless of who pays for them, and a new health intervention should be considered cost-effective if the expected outcome is valued more than the benefit foregone because society could not use the resources in their next best use. Of note, societies differ and many health economics studies are performed with a local national perspective. This should be kept in mind, as applicability to other societal milieus may be limited.

Time Horizon The investment a society is willing to make may vary depending on when the benefits are expected to be accrued. In addition, the available evidence may have limited value to perform accurate predictions for a long time horizon.

Discounting We give different value to the same item/occurrence depending on time, and that is called discounting. Discounting health benefits means that we prefer to see health benefits sooner rather than later. On the opposite, discounting costs means that we prefer to postpone paying for them. In evaluations with long term horizons a clear specification of discounting is warranted.

9.2 Commonly Used Outcome Measurements

Different metrics have been developed to assess the cost incurred to achieve a certain health gain. We will review the most commonly used ones.

DALY (disability-adjusted life year). DALYs are calculated adding the number of years of life lost due to premature mortality to the number of years of healthy life lost related to disability. One DALY is defined as the loss of the equivalent of 1 year of life at full health.

ICER (incremental cost-effectiveness ratio): it is calculated dividing the difference in total costs (incremental cost) by the difference in the chosen measure of health outcome or effect (incremental effect) to provide a ratio of extra cost per extra unit of health gain.

QALY (quality-adjusted life year): health states are assigned an arbitrary utility value from zero to 1; the incremental cost-effectiveness ratio per life year (ICER) is then divided by the utility. A year of life lived in optimal health is worth 1 QALY (1 Year of Life \times 1 Utility = 1 QALY).

WTP (willingness to pay): it is the amount a given society accepts to pay to gain 1 QALY. It is thought to be \$50,000 to \$100,000 in the USA and 20,000–30,000 pounds in the UK.

It should be noted that DALYs and QALYs are fundamentally different, and the exact same therapeutic effect may be appreciated differently depending on which one is used. Franco Sassi has shown that age of disease onset is an important factor to determine discrepancies. The QALYs gained tend to exceed the DALYs saved when disease starts in the very early years of life and is of short duration. When the disease starts in late adulthood and in older ages, the QALYs gained tend to exceed the DALYs saved. [Sassi F. *Calculating QALYs, comparing QALY and DALY calculations. Health Policy Plan.* 2006 Sep;21(5):402–8. <https://doi.org/10.1093/heapol/czl018>. Epub 2006 Jul 28. PMID: 16877455.]

9.3 Controversial Issues

Cost-utility studies are complex and require making assumptions that may be arbitrary, including very important value judgements that our patients may not share. Assigning value to a health benefit is an endeavor fraught with ethical challenges. Utility determinations that are made without public consultation can justifiably be perceived as a threat to patients' autonomy. Those challenges have been compounded by methodological issues.

Importantly, QALY's have been shown to assign lower value to the health gains of those who are disabled, elderly, and those who suffer a greater burden of disease. That happens because the quality of life of those with illness or disability is valued on the QALY scale lower than that of a healthier person. Those who are affluent and have better access to care are more likely to be healthier. Therefore, a public health decision made solely on QALYs may reinforce systemic biases and health inequalities, creating a "double jeopardy" for those at greater disadvantage.

In addition, we may assign different utility to the same intervention depending on what instrument we use to measure quality of life. Marra and colleagues examined the effect of using four different quality of life instruments to compare the cost-effectiveness of infliximab and methotrexate in rheumatoid arthritis. They found that, depending on the method used to measure quality of life and applied as utility to calculate QALYs, quite different incremental cost-utility ratios were generated. [Marra CA, Marion SA, Guh DP, Najafzadeh M, Wolfe F, Esdaile JM, Clarke AE, Gignac MA, Anis AH. Not all "quality-adjusted life years" are equal. *J Clin Epidemiol*. 2007 Jun;60(6):616–24. <https://doi.org/10.1016/j.jclinepi.2006.09.006>. Epub 2006 Dec 22. PMID: 17493521.]

Finally, QALYS have mathematical characteristics that make them unsuitable for the seemingly intuitive interpretation of its values in "year" units. QALYs are computed using variables measured on different scales. Life-years are expressed in a ratio scale with a true zero, whereas utility is an interval scale where 0 is the value assigned to death. [Prieto, L., Sacristán, J.A. *Problems and solutions in calculating quality-adjusted life years (QALYs)*. *Health Qual Life Outcomes* 1, 80 (2003). <https://doi.org/10.1186/1477-7525-1-80>].

9.4 Sponsorship Bias

We should keep in mind that studies of lesser methodological quality and/or funded by industry may be more likely to find that a new intervention or drug is cost-effective. In that sense, the sys-

tematic review by Bell and colleagues is quite relevant. They studied the characteristics associated with the probability that a cost-effectiveness study reported a favorable cost-effectiveness ratio. They found that studies funded by industry were more likely to report favorable ratios. In addition, studies of higher methodological quality and those conducted in Europe and the USA were less likely to report ratios below \$20,000/QALY. [Bell CM, et al. *Bias in published cost effectiveness studies: systematic review. BMJ.* 2006 Mar 25;332(7543):699–703. <https://doi.org/10.1136/bmj.38737.607558.80>. Epub 2006 Feb 22. PMID: 16495332; PMCID: PMC1410902.]

A subsequent large study by Xie and Zhou that assessed 8192 cost-effectiveness analyses found that among 5877 studies that reported positive incremental costs and quality adjusted life years, ICERs from industry sponsored studies were 33% lower (95% confidence interval – 40 to –26) than those from non-industry sponsored studies. [Xie F, Zhou T. *Industry sponsorship bias in cost effectiveness analysis: registry based analysis. BMJ.* 2022 Jun 22;377:e069573. <https://doi.org/10.1136/bmj-2021-069573>. PMID: 35732297; PMCID: PMC9214880.]

It seems prudent to avoid relying on industry sponsored cost-effectiveness studies when making public health policy decisions.

9.5 Role of Health Economics Studies in Healthcare Policy Making

When we decide to adopt or reject a new technology or drug, informative items to make that decision include the expected improvements in survival and quality of life, the cost incurred to achieve those improvements, the level of uncertainty regarding the cost effectiveness estimates, the burden of disease, the availability and cost-effectiveness of alternative treatments, the overall societal impact, and the expected effect on health inequalities and systemic bias.

The decision making process informed by a cost-effectiveness study should be carried out by a multi-disciplinary group that includes, or receives substantial input from, community represen-

tatives, public health experts, economists, statisticians, nurses, physicians, and healthcare administrators.

9.6 Checklist for Health Economics Studies

- Has the new treatment/technology been shown to be effective?
- Does the new treatment/technology compare favorably to the standard of care?
- Is the burden of disease in our society substantial?
- Does the study design consider and address existing health inequalities?
- Could there be sponsorship bias?
- Was quality of life measured in manner that is acceptable to my patients?
- What is the time horizon of the evaluation?
- Was discounting applied?
- Was the level of uncertainty in the measurements discussed?
- Are the results applicable to my patients? Will they inform beneficial change in my society?



10.1 Systematic Reviews and Meta-Analysis

It is hard for clinicians to remain up to date with the innumerable publications for each topic. Conflicting results (some findings are positive, while others are negative) make it even harder. A good systematic review offers a comprehensive view of the best available evidence for a given clinical question.

All meta-analyses include a systematic review, but not vice-versa. Sometimes the data are such that we cannot meta-analyze all individual results into a single pooled estimate. In those instances, we can still perform a systematic review.

Let us say we are interested in determining whether a new treatment is appropriate for a patient. The best-case scenario would be to find a meta-analysis that followed the PRISMA guidelines (<http://www.prisma-statement.org>), pooled data from good trials and reported clear benefit and low risk of adverse effects in patients similar to her. We would happily recommend such a treatment to our patient. The PRISMA guidelines are a minimum set of rules for the publication of systematic reviews and meta-analyses—akin to the CONSORT guidelines for clinical trials. The similarity to randomized trials does not stop there. We also require advance registration for systematic reviews and meta-analyses in a validated registry, such as PROSPERO. [<https://www.crd.york.ac.uk/prospero/>].

There is widespread consensus that a well-performed meta-analysis is at the very top of the evidence hierarchy in Medicine, but not all meta-analyses are equally good. Of note, if a meta-analysis was produced by the Cochrane Collaboration (www.cochrane.org) you can be certain that it was the result of methodical and thoughtful work. Their meta-analyses are the gold-standard in current literature.

We will review all the components needed to provide a good meta-analysis. At the end of the Chapter we will summarize them in a Checklist.

10.2 Publication Search

Meta-Analysis authors should use state-of-the-art search strategies in at least the following online databases: PubMed/Medline, clinicaltrials.gov, and the Cochrane Central Register of Controlled Trials. Technology has made it easier to replicate and verify publication searches.

When you read a meta-analysis produced outside of the Cochrane collaboration you may want to start at the end of the paper—that is where most journals tell you about funding sources. Beware of “obscure funding sources” for a meta-analysis, because they can be associated with selective (biased) publication searches. In some cases, the bias consists of Selective Reporting, which we will review later.

A clear flowchart, applying the PRISMA guidelines, should describe the search at the beginning of the manuscript. The example in Fig. 10.1 shows the work our group did searching for randomized clinical trials that assessed the efficacy of Community Health Worker interventions to improve glycemic control in minority patients with diabetes mellitus.

We began casting a very wide net that identified almost thirty-four thousand publications. In the screening phase, we narrowed down those results to Randomized Clinical Trials that had diabetes and/or glycemic control as disease of interest. We considered only full text articles as eligible and ended up finding only nine studies suitable for inclusion into the manuscript.

**Publication Search
PRISMA Flowchart**

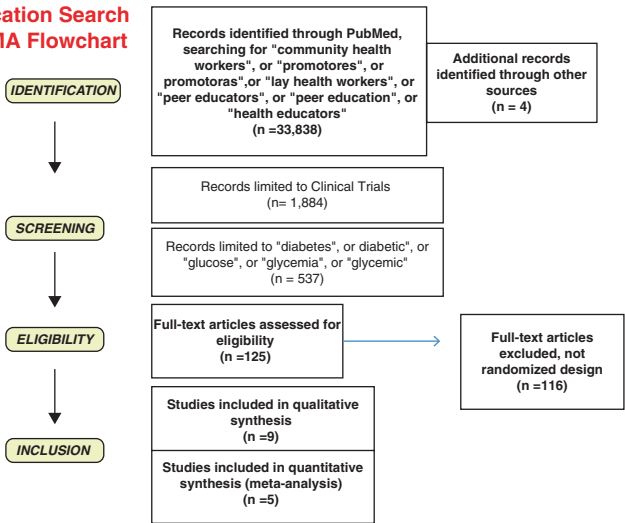


Fig. 10.1 PRISMA flowchart. The chart depicts the necessary steps to search for relevant studies

10.3 Publication Bias

Due to the preferential publication of positive studies, publication bias often results in an exaggerated estimate of the association or therapeutic effect of interest.

There is a predictable tendency for authors and editors not to publish results of negative studies (i.e., studies that fail to reject the null hypothesis). There is an even greater tendency not to publish negative studies if they were small, because the negative results could be attributable to lack of statistical power and lower precision.

10.3.1 Funnel Plot

We use the funnel plot as a visual tool to diagnose publication bias. It is a scatter plot that shows the relationship between the

effect size of each study and its precision. It usually depicts the Effect Size or Risk Estimate (such as the Relative Risk) on the horizontal axis and a measurement of the estimate's precision (such as $1/\text{Standard Error of the Relative Risk}$) on the vertical axis. As a result, we find larger studies with higher precision at the top of the graphic, while smaller studies with lower precision are located towards the bottom. If there is publication bias, smaller negative studies will be missing, creating asymmetry at the bottom of the graphic. Figure 10.2 shows an example of severe publication bias.

In Fig. 10.2 there are many studies of low precision that show a positive association (i.e. a log-odds ratio > 0), but none that show a negative association within that area of lower precision. There has been a bias against the publication of negative studies with lower precision. As previously discussed, smaller sample size is the most common cause of lower precision.

However, publication bias may not always be so overt, and we may need significance tests to come to our aid.

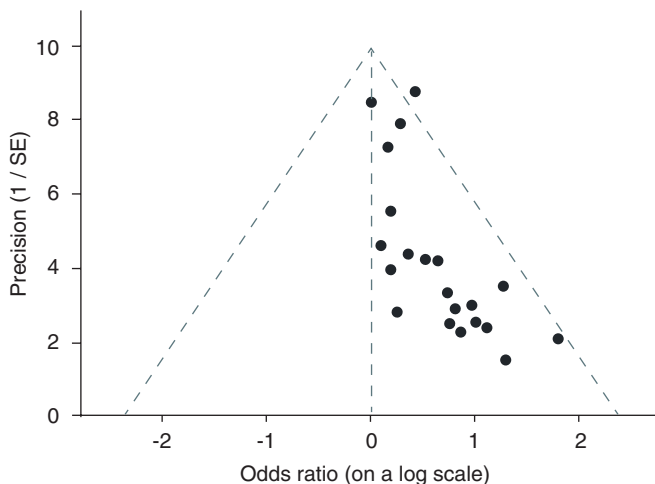


Fig. 10.2 Funnel plot with publication bias. The distribution of studies towards the bottom is truncated

10.3.2 Significance Tests

The visual assessment of a funnel plot to look for asymmetry can be difficult and quite subjective. Consider the funnel plots depicted in Fig. 10.3.

The original plot is on the left of the figure, and it clearly shows publication bias. However, it is not easy to decide when the bias is no longer there. Would it be enough to have three additional studies in the low precision-negative result area, as shown in the panel at the right? May be you need a few more—it is far from clear.

To deal with this type of issue several significance tests of publication bias have been developed. It is important to remember that the ability of those tests to detect publication bias is limited when the number of available studies is small (less than 10; see *Journal of Clinical Epidemiology Volume 53, Issue 11, November 2000, Pages 1119–1129*).

Let us consider Egger's regression test because it is a logical extension of the funnel plot.

Egger's test takes the funnel plot and keeps one of its axes unchanged: the one that describes precision as the reciprocal of the Standard Error ($1/SE$). The other axis uses the effect size from the funnel plot but divided over its Standard Error—we call that a Standardized Estimate of the effect size.

If there is no publication bias, the intercept of the regression line will be zero, or near zero. On the other hand, the test will be

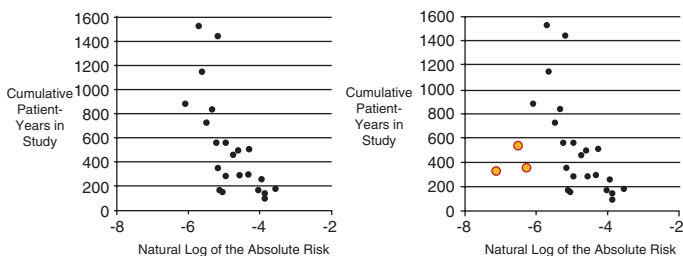


Fig. 10.3 Addressing publication bias. The right hand side shows the effect of adding three negative studies in an attempt to balance the lower part of the graphic

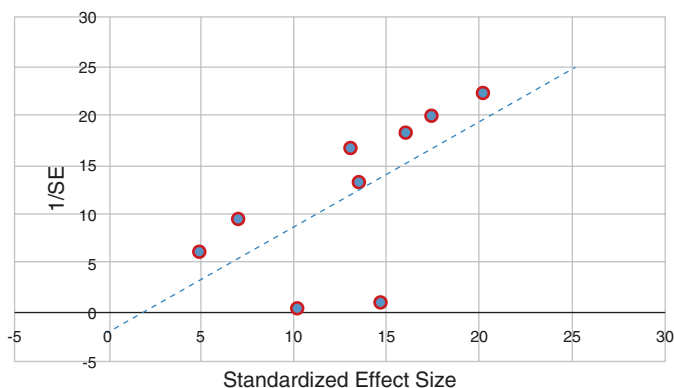


Fig. 10.4 Egger's regression test. The intercept is very close to zero, which strongly suggests there is no publication bias

significant for publication bias if the intercept differs significantly from zero. Figure 10.4 shows an example of an Egger's regression test when there is no evidence of publication bias. The intercept is very close to zero.

10.4 Selective Reporting

Selective reporting can be very difficult to detect and is thus very insidious. It may be unintentional, but it can also reflect an agenda that does not benefit society.

Unintentional selective reporting is just like Publication Bias, but within a single published report. Because nobody gets excited about results that are not statistically significant, sometimes authors will mention in passing that a certain event took place “at similar rates in both arms” without providing the readers with the actual data. Given that important adverse effects can be rare, a small incidence that failed to reach statistical significance in a single study may be significant in a meta-analysis. However, to perform a meta-analysis we need the actual data, not just a statement reassuring us that “it was not significant”. The solution to

unintentional selective reporting is already available. Virtually unlimited digital space online leaves no excuse to authors and editors because all “non-significant” data can be published as part of an online supplement.

Intentional selective reporting is quite more problematic and can have far reaching consequences, as we learned from the oseltamivir (Tamiflu) saga. Roche, the company that manufactures Tamiflu, held back important information from clinical trials until they were forced to release the data.

In 2009 there was great global concern about a possible influenza pandemic. Based on the available data there seemed to be several benefits from Tamiflu, including a positive effect to reduce the risk of complications such as pneumonia—a very attractive drug effect when you worry about a pandemic. As a result, countries and large corporations spent millions stockpiling the drug and turning it into a great moneymaker for Roche.

However, the overall impression of therapeutic benefit was driven by a combination of published trials and unpublished data given by Roche to Kaiser et al. (*Arch Intern Med* 2003; 163: 1667–72), which actually made their way into a Cochrane meta-analysis, biasing its result. The influence of the unpublished data regarding perceived benefits in the Cochrane Review caught the eye of Japanese pediatrician Keiji Hayashi, who in July of 2009 left an online comment about the strong influence of Roche unpublished data on the Cochrane findings. He suggested the unpublished data should be carefully scrutinized. His online comments were appreciated by the Cochrane collaboration, and they set in motion a process to obtain unbiased original trial data from Roche. The Cochrane collaboration, with help from the BMJ and the British media, exerted sustained pressure on Roche until the company finally released original trial data.

Meta-analysis of all trial data showed, for example, that oseltamivir modestly reduces the time to first alleviation of influenza symptoms, but it causes nausea and vomiting and increases the risk of headaches and renal and psychiatric syndromes (*BMJ* 2014;348:g2545). Hardly the kind of drug worth stockpiling.

10.5 Quality of Individual Studies—Risk of Bias

If the studies we include in a meta-analysis are of poor quality, we will have a great risk of getting biased meta-analytic findings. Therefore, it is essential to perform a systematic quality assessment of the studies we find. The Cochrane collaboration developed a user-friendly color code to summarize study quality. Each requirement to achieve good quality is a horizontal bar. Each study contributes equally to the bar, and its contribution is color-coded. Green means that the requirement was clearly met by the study. Yellow means that compliance with the requirement is unclear, or poorly described. Red means the study failed to meet the requirement.

In the example shown in Fig. 10.5 many of the individual studies (about 70% of them) were affected by incomplete outcome data. Selective reporting and lack of blinding were also problematic because they compromised more than a quarter of the studies. In this type of situation, we would not be confident about the quality of the available studies. We would consider the risk of bias for our meta-analytic findings to be substantial.

The components that are required in each RCT to ensure quality and reduce the risk of bias are the following:

1. Allocation concealment: randomization must be executed automatically, without human participation (e.g., you remotely

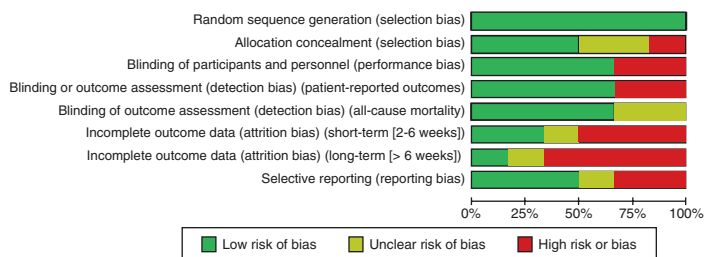


Fig. 10.5 Risk of bias graphic. The rows correspond to the items that determine the quality of each study. Poor quality increases the risk of bias

access a computer program to enter a unique participant ID, and the program assigns the randomization group). The process to allocate participants to treatment arms is thus concealed from the researchers.

2. Balanced distribution of baseline characteristics: this is pretty much guaranteed in very large trials with very simple randomization schemes, but may require stratified randomization, and/or blocks, in smaller studies.
3. Double blinding while the study was carried out.
4. Few losses to follow up.
5. Few crossovers.
6. Intervention fidelity: participants actually received the intended treatment (e.g., actual dosage, complete treatment period)
7. All relevant outcomes were reported, after being ascertained in a blinded, standardized fashion in most participants.
8. The results were analyzed applying the intention to treat principle.
9. A multivariate adjusted analysis was performed, if warranted.

10.6 Testing for Heterogeneity

We test for heterogeneity amongst individual study results as the final checkpoint before we pool our data into overall effect size estimates. We do it to ensure that we are not mixing up apples and oranges in our meta-analysis.

10.6.1 How Much Do Individual Study Results Differ from each Other?

The heterogeneity test examines whether the treatment effects are similar across studies. If we are pooling results about a dichotomous outcome, like mortality, it assesses whether the 2x2 tables summarizing each one of the studies are all originating from one large theoretical 2x2 table that represents the treatment effects in the population of interest (the blue table in Fig. 10.6). A signifi-

		INTERVENTION		CONTROL	
		122		200	
		INTERVENTION		CONTROL	
		142		110	
		864		803	
		INTERVENTION		CONTROL	
		202		100	
		804		902	
		INTERVENTION		CONTROL	
		202		100	
		803		902	
		INTERVENTION		CONTROL	
		222		140	
		784		862	

INTERVENTION	CONTROL
a	c
b	d

Fig. 10.6 Assessment of heterogeneity. A heterogeneity test assesses the probability that all study results correspond to an underlying shared distribution, depicted in blue

cant *P*-Value reflects significant heterogeneity—at least one of the individual studies seems to have a significantly different result.

Several test statistics may be reported for heterogeneity. However, I^2 is preferred, because it can be intuitively interpreted as the percentage of the total variability that is due to true heterogeneity, that is, beyond random variability.

I^2 values < 25, 25–50%, and 50–75% suggest low, moderate, and substantial heterogeneity, respectively.

$I^2 > 75\%$ suggests excessive heterogeneity → studies with such heterogeneity probably should not be meta-analyzed. A stratified analysis may be feasible.

When you read a meta-analysis look for the heterogeneity statistic in the graphic known as a Forest Plot. A conventionally formatted Forest Plot is shown in Fig. 10.7.

The point estimates of individual studies are depicted as boxes. The size of each box correlates to their sample size, and the corresponding lines are our old friends the Confidence Intervals. The diamond at the bottom represents the pooled effect size estimate and its confidence interval.

The I-squared statistic in this example is very high (92.2%). That is a clear indication of excessive heterogeneity, and it tells us

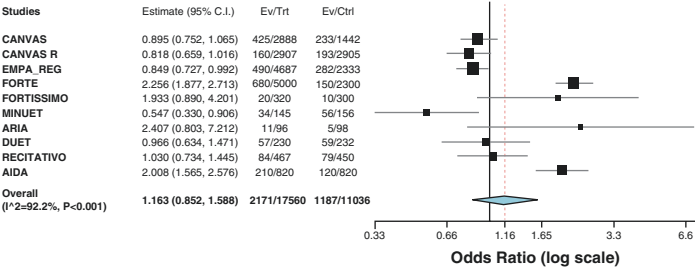


Fig. 10.7 Forest plot. The pooled estimate is depicted by the rhomboidal shape at the bottom

that we should not be obtaining a single pooled estimate from this dataset. Rather, we need to go back to the drawing board to identify the reasons for heterogeneity. That means re-examining how studies differed from each other regarding their design and methodology. We must look for clinically plausible explanations for heterogeneous results, and we must determine whether a stratified analysis would make sense.

10.7 Obtaining a Pooled Estimate

There are two main approaches to pool individual study results into a single meta-analytic statistic: The Fixed Effect and Random Effects models.

Those two methods give similar results when there is little heterogeneity among the studies, but they will give different results if there is substantial heterogeneity.

10.7.1 Fixed Effect Model

The Fixed Effect Model assumes that a single common (or ‘fixed’) effect underlies every study in the meta-analysis.

For example, we may be lucky enough to find many studies in which the same drug was used in every study, at the same dosage,

and in very similar patient populations. In that situation, if every study were infinitely large, they would all yield an identical result. The individual study results differ from each other due to random sampling variation but no more than that. Consequently, there would not be substantial heterogeneity and the I^2 should be $<50\%$.

The Fixed Effect method tends to result in narrower confidence intervals, and for that reason it has been inappropriately used when the Random Effects method should have been applied.

10.7.2 Random Effects Model

The Random Effects Model assumes that individual study results may represent slightly different treatment effects. For example, different drugs from the same class, but with varying efficacy, were used in each individual study.

The Random Effects approach takes that into consideration and allows for random error plus inter-study variability when calculating the pooled estimate and its confidence interval. It tends to result in more conservative effect size estimates with wider confidence intervals.

10.7.3 Fixed Versus Random Effects

If there is heterogeneity the confidence interval for the pooled effect size will be wider when the random-effects method is used and thus claims of statistical significance will be, appropriately, more conservative. The use of a fixed effect model requires low heterogeneity.

Let us see how in an example how the results we obtain can vary depending on the method. We are doing a meta-analysis of the value of intra-articular injections to relieve pain in severe osteoarthritis of the knee. We identify two types of studies: unblinded comparisons to usual care, and blinded comparisons to sham injection procedures. The outcome of interest was the relative risk of persistent severe pain at 3 months.

When we pool all the results into a meta-analytic estimate, we get different results depending on the method. First, let us see what happens when we use a Fixed Effect method (Fig. 10.8). It gives us a pooled Relative Risk (95% CI) for persistent pain of 0.906 (0.813 to 0.999). That is a 10% relative reduction in the risk of having pain at 3 months. However, there is excessive heterogeneity, with an I-squared of 79%.

If we apply a Random Effects method the pooled effect size is no longer significant, with a Relative Risk (95% CI) = 0.943 (0.738 to 1.15), as shown in Fig. 10.9. This is the correct approach.

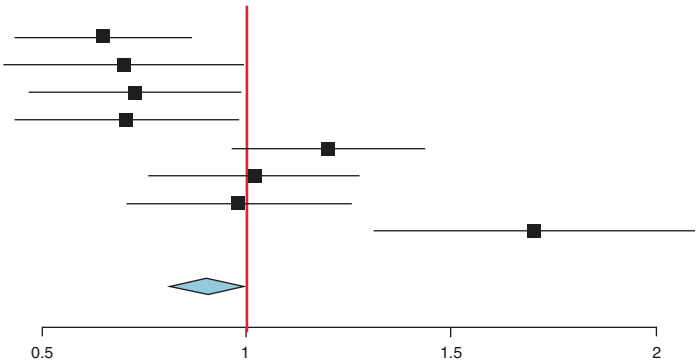


Fig. 10.8 Fixed effect method. The suggestion of efficacy by the pooled estimate (rhomboidal shape) may be spurious

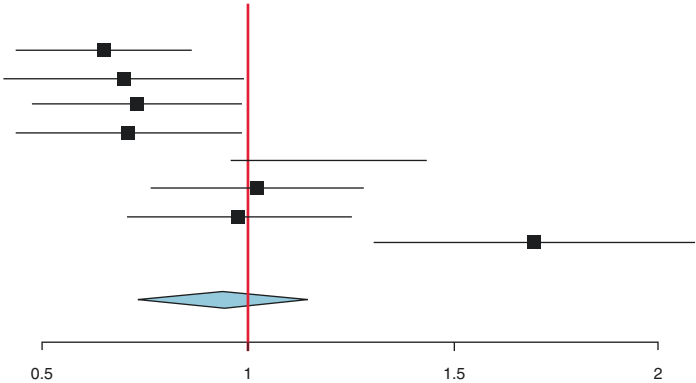


Fig. 10.9 Random effects method. The pooled estimate does not suggest efficacy

10.8 Sensitivity Analysis

In statistics performing a sensitivity analysis means changing one of our assumptions and repeating the analysis to examine the effect on the result. In meta-analysis, we perform sensitivity analyses to assess whether any given characteristic of the individual studies is associated with a difference in their outcomes. The study characteristic we assess through sensitivity analysis varies depending on the situation, and it could be quality, blinding, patient population, the drug used, the dosage used, the year of publication, etc. Let us go back to our example of intra-articular injections to relieve pain in osteoarthritis of the knee. We believe the results differed depending on whether the design was blinded or not, causing excessive heterogeneity. See below what happens when we stratify based on blinding.

Figure 10.10 shows the results of our sensitivity analysis. The vertical solid red lines in the graphic depict the null hypothesis value ($RR = 1$). The un-blinded studies suggested the presence of significant benefit (upper panel), whereas the blinded studies clearly

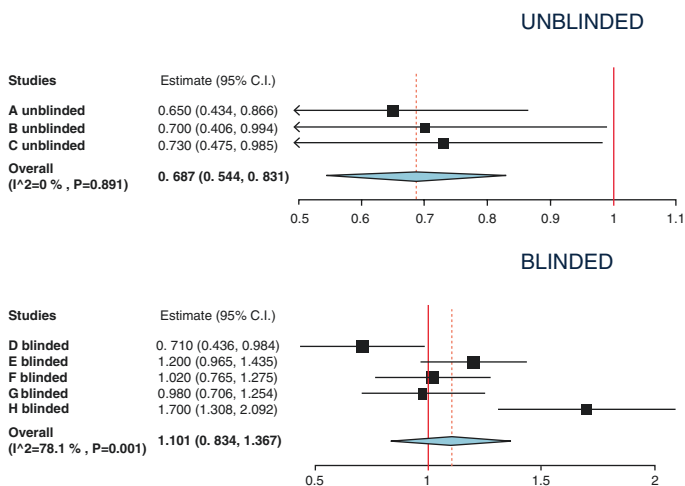


Fig. 10.10 Sensitivity analyses. Studies were stratified based on whether they were blinded

did not (lower panel). We value the information provided by blinded studies quite more because the outcome ascertainment is unbiased. We therefore conclude, based on the sensitivity analysis, that intra-articular injections probably will not help our patients.

10.9 Meta-Regression

Meta-regression is very similar to simple linear regressions. We want to determine how much the effect size estimate (Relative Risk, Mean Difference, etc.) varies depending on the level of a given variable. It is useful to examine whether the treatment effect, across studies, changes depending on that variable. This approach is more sensitive than stratified analysis to detect associations.

A great example of the value of meta-regression is the improved assessment of the consequences of vitamin E supplementation. Substantial controversy took place when some studies suggested that vitamin E supplements could be harmful. The meta-analysis by ER Miller et al. (*Ann Intern Med* 2005; 142:37–46) showed that there is an association between vitamin E and the all-cause mortality risk, but it is dose-dependent. Mortality risk is increased only by higher dosage, whereas lower doses are neutral or may even be beneficial.

Interestingly, if Miller et al. had limited their meta-analysis to a pooled estimate of vitamin E versus placebo, analyzing together the low-dose and high-dose studies without any stratification and without a meta-regression, they would not have had the same findings.

10.10 Network Meta-Analysis

Network Meta-Analyses create indirect comparisons between treatments when direct comparisons are scarce, but those treatments have been assessed against a shared comparison arm in other studies.

Network meta-analyses, like all other meta-analyses, should include a careful assessment of publication bias, quality of individual studies, and heterogeneity within pair-wise comparisons.

In addition, a network meta-analysis must have the following features:

- Similarity of patient populations across multiple studies. Indirect comparisons are accurate if the direct comparisons they are based on came from similar populations.
- Acceptable heterogeneity of the pairwise results pooled into indirect comparisons.
- Network consistency. All indirect comparisons should make sense in the context of the corresponding direct comparisons, and regarding other indirect findings.
- Acceptable degree of indirectness. If we only make indirect comparisons when contrasting two drugs, and there are no direct comparisons available, we would be much less certain about our findings.

There are two major aspects of a network meta-analysis that you should keep in mind because they determine validity of our model's network: one is assessed qualitatively, whereas the other one is a quantitative construct. They are Transitivity and Consistency. Transitivity is needed for all indirect comparisons, whereas coherence is a requisite in every loop of the network.

10.10.1 Qualitative Assessment of Transitivity

- All trials recruited similar patients for similar indications.
- All interventions were implemented in comparable fashion.
- Plausible effect modifiers are similarly distributed.

10.10.2 Quantitative Assessment of Consistency (a.k.a. Coherence)

It requires the availability at least some direct comparisons, so that the indirect comparisons may be contrasted against them. A test of significance is performed, and a small (significant) *P*-Value

indicates lack of consistency. Importantly, consistency is necessary for transitivity to be present. Substantial inconsistency compromises the strength of the network.

The methodology for network meta-analysis has evolved very rapidly and there is now validated software that performs sophisticated network meta-analysis, including Bayesian modelling. We will discuss an example of network meta-analysis generated with remarkable resource, available online, called MetaInsight [Owen, RK, Bradbury, N, Xin, Y, Cooper, N, Sutton, A. *MetaInsight: An interactive web-based tool for analyzing, interrogating, and visualizing network meta-analyses using R-shiny and netmeta. Res Syn Meth.* 2019; 10: 569–58]. I encourage you to learn more, visiting their site at <https://crsu.shinyapps.io/MetaInsight/>

We will use MetaInsight teaching data for a binary outcome, the success of smoking cessation interventions. To simplify, we will limit our models to trials assessing Individual Counselling, Group Counselling, and No-Contact.

Figure 10.11 depicts the strength of the evidence for each type of pairwise comparison. **In this kind of network graphic, the circles are called “nodes” and the connecting lines are “edges”.** The size of the nodes and the thickness of the edges in a network are proportional to the number of patients allocated to each intervention, and contributing to each one of the pairwise comparisons, respectively. In this case quite more data was available vis-à-vis the comparison of Individual Counselling to No-Contact.

Before we get to the Network part of the output, MetaInsight provides a graphic summary of the Odds Ratios for all available pairwise comparisons. This graphic is shown in Fig. 10.12, and allows a visual examination of trends in results distribution. It reinforces our assessment that there is an abundance of studies of Individual Counselling compared to No-Contact. Of note, an Odds Ratio > 1 in this setting is for a good outcome, success of the intervention for smoking cessation,

As discussed above, it is of fundamental importance to confirm there is Consistency between direct and indirect comparisons, so we are happy to see that the *P*-Values in Table 10.1 do not suggest significant inconsistency.

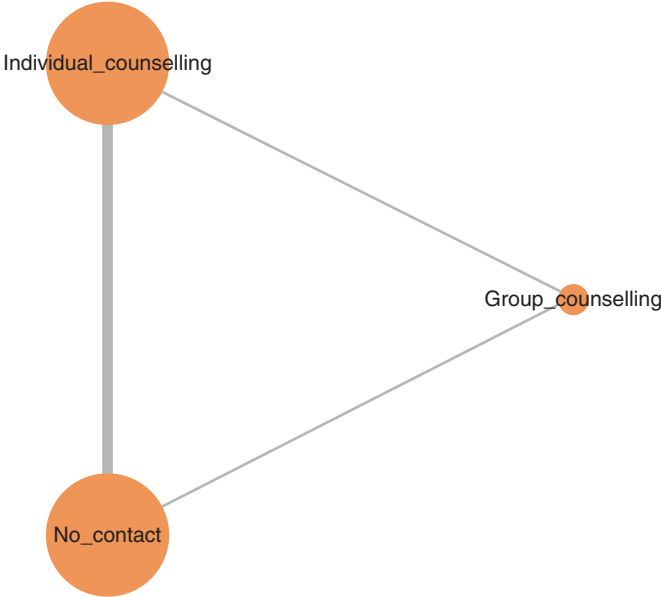


Fig. 10.11 Network meta-analysis. There are three nodes (treatments), connected by edges (comparisons)

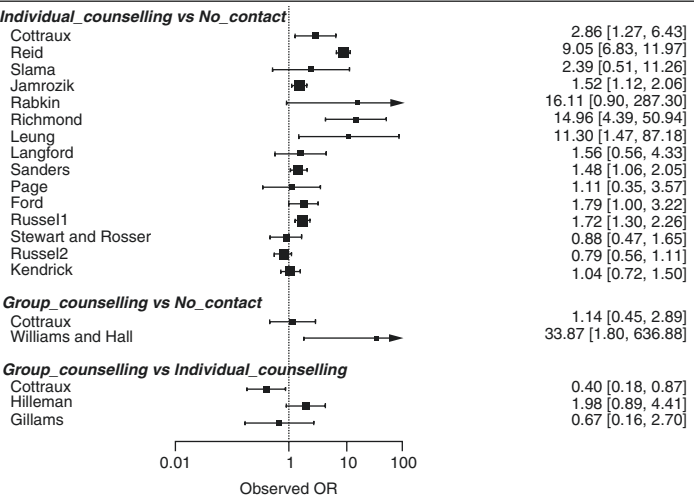


Fig. 10.12 Pairwise comparisons. Each pairwise comparison is akin to a conventional, non-network, meta-analysis

Table 10.1 Consistency between direct and indirect comparisons. There is no significant difference between the two types of comparisons

Comparison	NMA	Direct	Indirect	Difference	P-Value
Group_counselling:Individual_counselling	0.054996	-0.19866	1.378662	-1.57732	0.265511
Group_counselling:No_contact	0.832838	0.94761	0.750176	0.197434	0.858503
Individual_counselling:No_contact	0.777842	0.757183	1.667427	-0.91024	0.580689

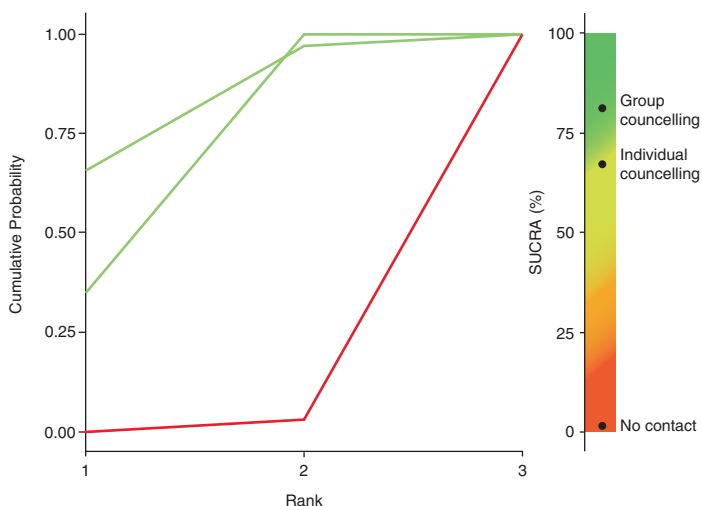


Fig. 10.13 Ranking of treatments. Group counselling ranks higher and has a larger area under the curve than Individual Counselling

A comparative ranking of network-assessed treatment efficacy is given by MetaInsight in both tabular and graphic format. The Bayesian graphics are extremely user-friendly, in spite of their nerdy names: Litmus Rank-O-Gram and Surface Under the Cumulative Ranking Curve (SUCRA). They make clear, because of the higher ranking location, and a curve nearer the top left, that Group Counselling had greater efficacy for smoking cessation in the network analyses, as compared to No-Contact (Fig. 10.13).

10.11 Meta-Analysis Checklist

- Was the meta-analysis registered a priori (PROSPERO, Cochrane) and compliant with PRISMA reporting guidelines?
- Was the publication search comprehensive?
- Was there evidence of publication bias?
- Was the quality of the individual studies appropriate?
- Was there evidence of selective reporting?

-
- Were the populations and interventions similar to my own?
 - Was there heterogeneity among selected studies?
 - What method was used to pool the results? Fixed or random effects? If a fixed effect method was used: was heterogeneity low?
 - What are the pooled estimates and their confidence intervals?
 - Were those estimates consistent across sensitivity analyses and in sub-group analyses?
 - Was a meta-regression performed?
 - What is the clinical significance of the results? Do they apply to my patient?
 - In a network meta-analysis, were the transitivity and coherence assumptions met?



Introduction to Artificial Intelligence Methods

11

11.1 Basic Definitions

Artificial Intelligence (AI) is the field of computing that aims at creating systems that mimic human cognitive abilities to solve tasks.

AI evolves rapidly and has an ever broadening scope, as better hardware and software become available. In Medicine, AI has been particularly apt at finding novel clinical solutions when provided with large volumes of data. That has already happened, for example, in the fields of Imaging, Genomics, Pathology, Oncology, and Cardiology.

Machine Learning (ML) is the specific part of AI that teaches a computer system how a given problem is solved through numerous examples, known as training set, with the goal of having the system solve a similar problem when exposed to it in the future.

Deep Machine Learning (DML) refers to more advanced machine learning that aims to create a system that has the ability to come up with its own solutions to the problem when it is given data and a clear definition of the desired outcome. It is called “deep” because it has a higher level of automation and because it has more than three layers of neurons.

Neural Networks are the backbone of machine learning. Neural networks are meant to mimic human brains and are composed of neurons (also called nodes), which are set up in layers.

The first layer takes the information in and is known as the **input layer**. The final layer is the **output layer**. In between we have the **hidden layers**.

Networks use three main components to produce an output: inputs, weights, and thresholds.

For example, if a neural network is created to diagnose Atrial Fibrillation the input will be electrocardiographic data, which will be deconstructed into components called features. Each feature will be given a weight. Each layer of neurons will produce an output based on features and their weights, in a process similar to producing regression probability estimates. If the output of any individual node is above a specified threshold value, that node is activated, sending data to the next layer of the network. However, the outputs will be sometimes wrong in diagnosing Atrial Fibrillation—there will be an error rate. The process of fine tuning the weights we give to our features to reduce the error rate observed in the preceding layer is what is known as **back-propagation**. Back-propagation is the basic learning mechanism of a neural network and is iterative in nature.

Of note, there are two definitions in AI that may be confusing to clinicians:

1. Predictive variables in the input are called “features”.
2. The threshold we described above as the value that activates the node is also called “bias”. The use of the word “bias” in this manner is unfortunate because bias, in the way we define it in evidence-based medicine, can be an important problem in AI applications, as we discuss below.

When training a network, the difference between the predicted values of the outcome and the actual values is known as a **Loss Function**. The procedure that systematically adjusts the weights and thresholds to minimize loss is known as the **Optimizer Algorithm**. Just like in old fashioned regression models we want to minimize our errors but without **overfitting** to the training dataset, because overfitting can reduce reproducibility of the model in future datasets. In other words, the network should not be so finely tuned to the training data that it has difficulty classifying other

similar data in the future. To reduce the risk of overfitting we are parsimonious in selecting features for our model, we try to stop the iterations as soon as the model fits well enough, and we need to be sure our training data set is large enough.

Supervised Vs. Unsupervised Learning In supervised learning we tell the system what variables to use, and what the desired outcome is. In machine learning lingo we call that using Labelled Input and Labelled Output. In unsupervised learning we give the system raw data and we let it “figure out” by itself what the meaningful features and patterns are. A seasoned biomedical researcher will probably view unsupervised training with great caution, and will probably consider its results exploratory until plausible biological mechanisms and reproducible clinical outcomes provide consistent validation of said results.

Convolutional Neural Networks are used to process visual data, such as images and video. They are named after convolution, which is the process that takes an image and extracts from it all the relevant features such as edges, color, gradient, etc., and then uses those features to classify the image. A convolutional network learns by itself the features of an image that are most informative to achieve accurate classification, rather than being told which features to use.

11.2 Performance Analysis of Machine Learning Models

After developing a model in the training data researchers should check the model’s performance in an independent validation data set that represents our population of interest as much as possible.

If the outcome of interest is dichotomous, such as disease/outcome being present vs. absent, a Receiver Operating Characteristic Curve (ROC) analysis should be provided. This ROC analysis is the same we have discussed in the Assessment of Diagnostic Tests chapter. It is a robust test to assess the ability the model has to tell apart those with the outcome from those without it, regardless of the proportion with the outcome in the population. In other words,

ROC curves can reliably assess the discriminant accuracy of models regardless of the disease prevalence. This is important because some researchers report the F test as a measurement of model accuracy in AI. As we discuss in the pertinent chapter, F -test values have a positive correlation with outcome prevalence, and should thus be interpreted with caution.

11.3 Biases in Artificial Intelligence

The potentially revolutionary benefits we may accrue from AI should not make us forget that, like any other endeavor, AI may be compromised by bias and errors. There have been several examples of bias in AI that were identified only after their effects on people had already taken place. We will discuss a few sobering examples next.

In 2018 the news agency Reuters reported the trouble created by an AI recruitment tool at Amazon. The company needed to streamline their resume review process. A computer engineering team reportedly created 500 computer models focused on specific job functions and locations. They taught each to recognize some 50,000 terms that showed up on past candidates' resumes. That training data set was heavily male. The system "learned" to ignore basic traits, such as code writing skills, because they were shared by most applicants. Instead, it selected terms that favored male candidates. This is an example of how a bias (in this case gender bias) can start at the training data set and end up compromising the output of a machine learning algorithm. [<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon--scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>].

Healthcare systems often use commercially developed algorithms to inform policies that affect millions of patients. Obermeyer et al. reported in 2019 that a widely used algorithm used to identify patients who would benefit from a referral to a program with more personalized care favored white patients. Equally sick black patients were less likely to be referred to those programs. The reason for racial bias was the use of health care expenditures as a proxy for higher risk prediction. White patients had historically better access to expensive care and therefore their

expenditures were higher at the same severity of sickness. The algorithm took the biased information from the training dataset and erroneously predicted white patients to be at higher risk, giving them prioritized access to more personalized care. AI reinforced and exacerbated systemic racism. [Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. *Science* 336, 447–453 (2019)].

Another striking example of systemic racism being amplified by an algorithm was COMPAS (Correctional Offender Management Profiling for Alternative Sanctions). COMPAS was an algorithm used in US court systems to predict the probability a defendant would become a recidivist. Dressel and Farid showed that the model predicted twice as many false positives for recidivism for black offenders (45%) than white offenders (23%). In addition, COMPAS was not more accurate or fair than predictions made by people with little or no criminal justice expertise, and a simple linear predictor provided with only two predictive variables was nearly equivalent to COMPAS with its 137 predictive variables [Dressel J, Farid H. *The accuracy, fairness, and limits of predicting recidivism*. *Sci Adv*. 2018 Jan 17;4(1):eaao5580. <https://doi.org/10.1126/sciadv.aao5580>. PMID: 29376122; PMCID: PMC5777393.]

Bias may arise in AI algorithms at each step: training data selection, code programming, and/or outcome definition. Societal and systemic biases have been reinforced and amplified by AI. AI-specific safeguards against bias are warranted.

11.4 Quality of Medical Research Using Artificial Intelligence

Nagendran and colleagues performed a systematic review of studies comparing the performance of diagnostic deep learning algorithms for medical imaging with that of expert clinicians. They found only 10 records randomized clinical trials, two of which had been published. Those two suffered from lack of blinding.

There were 81 non-randomized clinical trials, but only nine were prospective and just six were tested in a real world clinical setting. Full access to all datasets and code was severely limited (unavailable in 95% and 93% of studies, respectively). The overall risk of bias was high in 58 of 81 studies and adherence to reporting standards was suboptimal (<50% adherence for 12 of 29 TRIPOD items). 61 of 81 studies stated in their abstract that performance of artificial intelligence was at least comparable to (or better than) that of clinicians. Nagendran et al. recommended that “future studies should diminish risk of bias, enhance real world clinical relevance, improve reporting and transparency, and appropriately temper conclusions.” [Nagendran M, et al. *Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies* BMJ 2020; 368: m689 <https://doi.org/10.1136/bmj.m689>].

The study by Nagendran et al. highlights the fact that AI-based research must be subject to the same rigorous examination that we apply in other areas of Evidence Based Medicine. In that sense, it is encouraging to see the efforts to develop a reporting guideline (TRIPOD-AI) and a risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. [Collins GS, Dhiman P, Andaur Navarro CL, et al. *Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence*. BMJ Open 2021;11:e048008. <https://doi.org/10.1136/bmjopen-2020-048008>].

11.5 Checklist for Artificial Intelligence Methods

- Does the study comply with evidence-based guidelines applicable to its basic design and goals: study of etiology, diagnosis, screening, prediction, or randomized clinical trial?
- Are the training and validation data sets representative samples of our population of interest?
- Is the training data set large enough?

- Has the possibility of systemic and/or societal bias been explored in the data? Racial and gender bias should be specifically addressed.
- If the data were retrospectively acquired, were known biases examined?
- How are missing data treated? Are there sensitivity analyses using different methods?
- Did the researchers use supervised or unsupervised learning?
- Was the optimized model clearly described? Do the selected features (predictive variables) have a biologically plausible role in your clinical model of disease?
- How was performance of the model in the validation set evaluated? Is a Receiver Operating Characteristic Curve analysis provided for dichotomous outcomes? If an F-score test is provided, is the outcome prevalence similar to that in your population of interest?
- Is the selected outcome clinically relevant? Does it have societal value? Is it free from bias?
- Does the model achieve a clinically relevant outcome in prospective real-world validation?



12.1 The Essential Components of Clinical Practice

Each step in our clinical work should be based on the best available evidence, as follows:

Etiology and Epidemiology: we need to understand the causes of a disease process and the behavior of that disease process in our population.

History and physical examination: all items should be validated to optimize our pre-test probability assessment.

Diagnosis: the most accurate tests should be selected, and applied based on pre-test probability estimates. We prioritize the diagnosis of the more severe and treatable processes.

Prognosis: we aim to predict the most likely clinical course, in order to inform the patient and their family.

Therapy: we preferentially use the most efficacious, acceptable and cost-effective treatments.

Shared Decision Making: treatment is a collaborative process based on the patient's needs, expectations, and desires.

Prevention: we prevent disease through reduction of risk factors, and we reduce morbidity and mortality through screening and earlier treatment.

12.2 Five Step EBM Model

As described by David L. Sackett in 1997 the five steps in the evidence-based model of clinical practice are:

1. Formulate an answerable clinical question. Basic components of an answerable question: **PICO** (**P**atient, **I**ntervention, **C**omparison, **O**utcome).
2. Find the evidence: Cochrane Library, MEDLINE, trusted EBM sites.
3. Appraise the evidence: evaluate the quality of studies, using design-specific checklists (meta-analysis, randomized controlled trials, studies of diagnostic tests, observational studies).
4. Apply the evidence: make collaborative diagnostic and treatment decisions.
5. Evaluate the performance: assess outcomes at patient and population levels.

[Sackett DL. Evidence-based medicine. *Semin Perinatol.* 1997 Feb;21(1):3–5. [https://doi.org/10.1016/s0146-0005\(97\)80013-4](https://doi.org/10.1016/s0146-0005(97)80013-4). PMID: 9190027.]

12.3 The Hierarchy of Evidence

Not all types of studies have equal value. At the very bottom, providing the lowest value, are Opinions, Case Reports and Case Series. After that we have Case-Control and Retrospective Studies. Next come Prospective Observational Studies. At the very top we have Randomized Clinical Trials and Meta-Analysis.

A methodologically sound Meta-Analysis that includes several good quality Randomized Trials is generally viewed as the strongest type of evidence in Medicine.

12.4 The Cochrane Collaboration

Your evidence search should probably begin at the Cochrane collaboration site: <http://www.cochranelibrary.com>. The Cochrane collaboration is the world leader in the development and implementation of the best meta-analyses to-date. There are more than fifty Cochrane Review Groups (CRGs), responsible for preparing and maintaining reviews within specific areas of health care.

Once within the Cochrane Reviews page, my preference is to click on the Clinical Answers tab, which can be queried by Topics. If your question is not addressed in Clinical Answers, you should proceed to the Special Collections tab, followed by the Cochrane Reviews tab.

Although the Cochrane collaboration has rightfully gained a reputation as the gold standard in meta-analysis, it is not infallible. For example, its HPV vaccine review has been reported as flawed (Jørgensen L, Gøtzsche PC, Jefferson T. *BMJ Evidence-Based Medicine* 2020;23:165–168).

12.5 PubMed Queries

A good starting point in PubMed is provided by their Clinical Query Tool, which can be found at <https://pubmed.ncbi.nlm.nih.gov/clinical/>. There you will see a filter that allows you to select studies pertaining to: Etiology, Clinical Prediction Guides, Diagnosis, Therapy, or Prognosis.

You should enter your clinical question in the simplest possible terms, such as “SGLT-2 inhibitors for the treatment of CKD”. When the results are displayed, it is a good idea to click on the link that appears right on top of the study list: “See all results in PubMed”. Once in that PubMed results window you should consider first clicking on the box that selects the “Article Type” as Meta-Analysis. In the ideal scenario you will find a Meta-Analysis published in a reputable journal, such as the New England Journal of Medicine, JAMA, BMJ, or The Lancet.

After examining the available meta-analysis, you should go back to the first PubMed results window and change the filter to “Randomized Controlled Trials”. I find it useful to also click on the pull down box “Display options” and select “Best match”.

12.6 Other Reliable Sources

In addition to the Cochrane Collaboration and PubMed you should consult other reliable sources on a regular basis. Those sources include:

- The ACP journal club at <http://www.acponline.org/journals/acpj/jcmenu.htm>
- The National UK Health Service Centre for Reviews and Dissemination at <http://nhscrd.york.ac.uk/welcome.html>
- Bandolier at <http://www.jr2.ox.ac.uk/Bandolier>

12.7 Google Scholar

Google Scholar is a frequently used search tool because it prioritizes relevant articles and provides links to publications that can be accessed in full-text format and free of charge. Many publications listed in PubMed cannot be fully accessed for free. A study by Salimah Shariiff and colleagues found that for quick clinical searches, Google Scholar returns twice as many relevant articles as PubMed and provides greater access to free full-text articles. [Shariff SZ, et al. *Retrieving clinical evidence: a comparison of PubMed and Google Scholar for quick clinical searches. J Med Internet Res.* 2013 Aug 15;15(8):e164. <https://doi.org/10.2196/jmir.2624>. PMID: 23948488; PMCID: PMC3757915.] It should be noted, however, that the quality of the available full-text publications was not assessed. In addition, their study was meant to identify features applicable to quick searches, not to comprehensive queries routinely used by meta-analysis groups, such as the Cochrane collaboration. An up-to-date Cochrane review remains

far superior to any quick search through Google Scholar. Finally, the qualities of Google Scholar should not be attributed to the main Google engine, which remains susceptible to strategies that deliberately enhance the visibility of certain citations/publications.

12.8 The Future of EBM

The future of EBM is in your hands. To maximize the positive impact of EBM we should collectively continue to:

1. Identify areas where evidence is lacking or incomplete.
2. Identify and address systemic biases.
3. Ensure an ethical approach to future research.
4. Educate society about the benefits of an evidence based approach to the practice of medicine.
5. Actively engage in community and institutional activities to promote best clinical practices. This must include governmental and political activities. Policy is too important to be left to politicians.
6. Promote awareness among clinicians. Our patient comes to see their doctor, not a hospital employee, or a mere dispenser of the latest fashionable untested treatment.



13.1 Relevance

There have been horrendous episodes in history, in which physicians actively participated in horrendous crimes against humanity. Racism, and the worst kind of paternalism, have also permeated “research studies” in not so distant history, such as the Tuskegee study, which continued until the 1970’s.

People from lower income countries have been recruited more frequently recruited into higher risk clinical research. Lower income people from our own society may participate, in disproportionate numbers, in higher risk studies.

In addition, people may be invited to participate in relatively risky studies with a complicated statistical design, that their own physicians do not fully understand.

Although bioethical principles may seem self-evident, they have been spelled out and refined over time. They should also evolve as our social and medical circumstances evolve. They should protect all people, regardless of race, ethnicity, gender, or sexual orientation. Some basic principles may sometimes clash with each other, and a careful conversation is always warranted.

13.2 The Belmont Report

Starting in the 1930's, the Tuskegee syphilis study in Alabama used disadvantaged, rural black men to study the "natural history" of the disease. Black men known to have syphilis were not informed of their status, nor were they offered penicillin after it had already been shown to be an effective cure for the disease. The study continued, under numerous US Public Health Service supervisors, until 1972, when a leak to the press resulted in its termination. The victims of the study also included 40 wives who contracted the disease, and 19 children born with congenital syphilis.

The atrocious Tuskegee study led to the Belmont Report, in 1979. The three basic principles outline in the Report remain a most useful tool to examine clinical research from an ethical perspective. Those principles are: Beneficence, Justice, and Respect for Autonomy.

13.3 Beneficence

Beneficence means that we should always strive to do what is best for the research participant. This is an extension of the well-known principle of "First Do No Harm". When designing research studies, we should maximize possible benefits and minimize possible harms to participants.

An important implication of this principle is that whenever we evaluate a new treatment the control arm of the study must represent the best available standard of care for that condition. Clinicians should not accept as ethical a study control treatment that they would not recommend as "the best available" to their own patients. For example, some studies have used drugs that require twice or thrice daily dosing in the control arm when a similar once daily formulation was already available.

The Benevolence principle requires a careful and nuanced evaluation when the research evaluates a higher risk intervention aimed at reducing very severe outcomes. Technological advances have led to increasing numbers of such "high risk, but high yield" interventions in Medicine.

13.4 Justice

Participants in a trial should stand to benefit, as a group, from the results of the study. When we say “as a group”, it means that actual participants may never benefit, but similar patients would.

For example, it is unethical to perform research about a new, expensive, drug in a lower income country if the manufacturer intends to market the drug in ways that make it accessible only to affluent populations.

13.5 Respect for Autonomy

Autonomy is a fundamental human right. In research, Autonomy means that participants must willingly enroll in a study without any coercion, whatsoever, and that they must be fully informed about the risks implicit in their participation. In addition to being aware of all potential risks, they should also understand who could benefit, and how.

Not everybody is capable of making their own decisions (children, people without capacity), and they should be adequately protected when others are making decisions for them.

The Informed Consent process is essential to inform the participant and protect their autonomy, and that is why the work of Institutional Review Boards is so important.

In addition, there is a very sensitive area that always requires close examination: financial compensation of participants. There is a clear need for participants to be compensated for transportation, meals, lost wages, etc. However, some higher risk research is compensated in a manner that makes financial benefits the preeminent reason to participate. This has generated very troubling situations such as when clinical research centers that recruit into handsomely paid studies are located very close to homeless shelters.

13.6 Institutional Review Boards

Institutional Review Boards (IRBs) perform an extremely important role in protecting the rights of research participants. The IRB

must ensure that the study addresses a significant scientific question, which has value for society at large, and that the societal benefit could not be achieved without conducting the study. Study design must be adequate to ensure that meaningful results will be obtained at completion.

The IRB must consider at least the following questions when assessing a proposal:

- Is the proposed study scientifically sound? Is it expected to provide meaningful information?
- Are all risks to participants acceptable and minimized?
- Are study subjects selected in an equitable fashion that reflects how the results will be used?
- Are there adequate provisions to protect privacy and confidentiality of study subjects?
- If vulnerable populations will be studied, are adequate protections in place to safeguard them?
- Will valid informed consent be obtained?
- Is the Consent Form optimally written, and does it address the specific needs of the population being recruited into the study?

IRBs constantly face multiple challenges, including the difficulty recruiting and retaining qualified individuals who have the requisite scientific expertise. In addition, it is essential to actively avoid all conflicts of interest.

13.7 Conflicts of Interest

Individual conflicts of interest have been egregious in the past, and have led to ongoing monitoring by academic medical centers of their faculties' potential conflicts.

Institutional conflict of interest may be much subtler and thus harder to prevent. A university or academic medical center benefits financially and gains prestige when it conducts high profile research. In addition, IRB members may feel pressed to approve protocols if the researcher is a powerful person in their institution.

There may also be a tendency to protect the institution's reputation when dealing with adverse events or possibly unethical behavior. In spite of all the challenges, IRBs perform extremely important work, and they need your active involvement to better protect research participants.

Appendix: Self-Assessment Test

1. A randomized controlled trial evaluated lidocaine patches, as compared to placebo patches, for back pain. Response to treatment was evaluated through a questionnaire, obtained at baseline and at 6 weeks. Thirty-five percent of participants randomized to lidocaine patches and 33% of those randomized to placebo patches did not complete the 6-week questionnaire. Which one of the following was a major concern when analyzing the data? Please select the one TRUE answer.
 - A. High probability of effect modification.
 - B. High probability of type 1 error.
 - C. High probability of type 2 error.
 - D. High probability of effect medication.
 - E. High probability of confounding.
2. Which of the following are fundamental ethical principles governing the conduct of clinical trials?
Please select the one TRUE answer.
 - A. The question is relevant, and cannot be answered through other means.
 - B. Participation is voluntary, after informed consent is provided.
 - C. The study addresses a reasonable doubt regarding the effectiveness of the treatment being assessed—there is “equipoise”.

- D. All of the above are correct.
E. A and B are correct, but not C.
3. You are very excited about a new randomized trial showing that dapagliflozin reduces mortality in patients with heart failure who have an ejection fraction $< 40\%$. The study was very well-designed, double-blinded, and placebo-controlled. It reported that at a median follow-up of 18 months, death from any cause occurred in 276 of 2373 patients randomized to dapagliflozin and 329 of 2371 patients randomized to placebo. Which one of the following statements is FALSE? Please choose one.
- A. The estimated relative risk is 0.84.
B. The estimated relative risk reduction is 0.16.
C. The estimated number needed to treat is 44.
D. The estimated absolute risk difference is 0.16.
E. The estimated event rate in the treatment arm is 11.6%.
4. You are seeing patients in the Emergency Room after a 2-week elective rotation. Your first patient is a 54-year-old male patient, with history of type-2 diabetes mellitus, hypertension, and hyperlipidemia. He presented with non-specific chest pain. His ECG is unchanged from one obtained 1 year ago.

You look up his pre-test probability of his chest pain being due to Coronary Artery Disease. A validated scoring system gives you an estimate of 20%.

You order a highly sensitive Troponin I test which has a reported Sensitivity of 98% and Specificity of 90% at a pre-determined cutoff value. The result is higher than that cutoff value.

Which one of the following statements is FALSE? Please select one.

- A. The Likelihood Ratio for the positive result is 9.8.
B. The prevalence of CAD, before a Troponin test, is estimated as 20%.
C. The pre-test odds of CAD are 0.25.
D. The post-test odds of CAD are 0.20.
E. The Likelihood Ratio for a negative Troponin test would be 0.02.

5. You decide to complete a 2x2 table with the data corresponding to the same patient (see below).

Based on this table, which statement is FALSE? Please select one.

	Disease	No Disease
Test Pos.	196	80
Test Neg.	4	720

- A. The Likelihood Ratio for the positive result is 9.8.
B. The Negative Predictive Value would have been 0.29.
C. The Positive Predictive Value is 0.71.
D. The pre-test probability of CAD, before a Troponin test, is estimated as 20%.
E. The Likelihood Ratio for a negative result would have been 0.02.
6. If you used the same Troponin test with the same cutoff for diagnosis, but in a patient with an estimated 60% pre-test probability, which one of the following statements would be TRUE? Please choose one.
- A. The Negative Predictive Value would go down
B. The Specificity would go down.
C. The Sensitivity would go down.
D. The Positive Predictive Value would go down.
E. None of the above.
7. Your medical center is interested in the potential efficacy of hydroxychloroquine to reduce mortality in patients with COVID-19 that required hospital admission due to hypoxemia. You find a meta-analysis of 3 observational studies. In addition, there was a separate large randomized clinical trial. All the studies recruited your population of interest. The pooled efficacy estimate in the meta-analysis was a Relative Risk (95% Confidence Interval) of 0.89 (0.79 to 0.99). The randomized trial reported a Relative Risk (95% Credible Interval) of 0.99 (0.87 to 1.11). What should you tell your hospital administrators? Please select the one TRUE statement.

- A. The best available evidence is provided by the meta-analysis.
 - B. The best available evidence is provided by the randomized trial.
 - C. The confidence interval is more reliable than the credible interval.
 - D. A and C are true.
 - E. B and C are true.
8. A randomized clinical trial finds that a new drug to treat Generalized Anxiety Disorder is more efficacious than the standard of care drug. It reduces the incidence of anxiety episodes by 27%. The P-Value for the comparison is 0.001.
- What does this mean? Please select the one **TRUE** answer.
- A. There is a 0.001 probability that the new drug is worse than the standard of care.
 - B. There is a 0.001 probability that this difference in efficacy, or a larger one, would be found if the null hypothesis was true.
 - C. There is a 0.001 probability that the standard of care is better.
 - D. There is a probability of 0.001 that this difference in efficacy would be found purely due to random error.
 - E. There is a 0.001 probability of type 2 error.
9. A third variable that is independently associated with both the exposure and the outcome, does not mediate the relationship between the exposure and the outcome, and should be considered when analyzing the association between exposure and outcome is known as (please choose one **TRUE** answer):
- A. A confounder.
 - B. An effect modifier.
 - C. A dichotomous variable.
 - D. A dependent variable.
 - E. An ordinal variable.
10. Which one of the following is **not** required to ensure adequate statistical power of a randomized clinical trial? Please select one answer.
- A. A low number of cross-overs.
 - B. A high follow-up rate.

- C. A large enough sample size.
 - D. An accurate prediction of the effect size.
 - E. Absence of protopathic bias.
11. A PET-CT method has been widely used to diagnose a type of cancer for about 5 years. It is regarded as the best available test for that condition. A group of investigators in your institution decides to use a large clinical database to assess the accuracy of PET-CT to diagnose that cancer. They propose to include in the study all patients who underwent PET-CT for that indication over the last 3 years, and who had pathologic confirmation in a biopsy as the gold standard.

What is your concern about the proposed study design?

Please select the one TRUE answer.

- A. The possibility of spectrum bias.
 - B. The possibility of interviewer bias.
 - C. The possibility of post-test referral bias.
 - D. The possibility of recall bias.
 - E. The possibility of prescription bias.
12. What type of design prevents post-test referral bias when assessing the accuracy of diagnostic test? Please select the one TRUE answer.
- A. Prospective design with all participants undergoing the test and the gold standard.
 - B. Retrospective design with all participants undergoing the test and the gold standard.
 - C. Case-control design, sampled based on gold standard availability.
 - D. Nested case-control design, sampled based on convenience.
 - E. Open label design.
13. Your institution intends to purchase a laboratory kit to measure N-terminal B-type natriuretic peptide (NT-BNP) levels to diagnose congestive heart failure. There are 2 manufacturers who offer very similar methods, but at a very different price. There has been one large, well-designed study comparing the diagnostic accuracy of those two methods head-to-head. It reported that method A (the more expensive one) had an area under the curve for the ROC of 0.92. Method B had an

AUC of 0.93. The P-Value for the comparison was 0.25, and the estimated statistical power was 92%.

What advice would you give to your hospital administrators? Please select one answer.

- A. They should probably purchase kit B.
 - B. They should probably purchase kit A.
 - C. They should probably purchase either kit, as they are similar.
 - D. They should wait for further studies comparing the 2 kits before making a decision.
 - E. They should look for other kits in the market because the accuracy of these 2 was too low.
14. The situation in which a third variable significantly modifies the association of the exposure with the outcome is known as (please select the one TRUE answer):
- A. Confounding
 - B. Interaction
 - C. Stratification
 - D. Bias
 - E. Unmasking
15. You are admitting an 81-year-old lady who presents with acute onset of shortness of breath at rest, which started 2 hours ago. She underwent right hip replacement 10 days ago at another hospital. Her only prior medical history is type 2 diabetes, which is well controlled with diet only. She denies taking any prescription medications.

At the exam she is tachycardic (110 beats per minute) and remains short of breath, but there absolutely no exam signs of congestive heart failure. You find a palpable tender cord at her right calf. Her ECG is completely normal. After a quick chat with your attending you ascertain her score using the Wells prediction rule for pulmonary embolism (PE). It is 9 points. Her bleeding risk score is low for her age.

Which one of the following statements is FALSE? Please select one statement.

- A. Her estimated risk of having PE is high.
- B. Other diagnoses, alternative to PE, are unlikely.
- C. Her estimated risk of having PE is about 78-80%.

- D. Anticoagulation should be deferred until after a pulmonary angiogram confirms the diagnosis.
 - E. Anticoagulation should be offered immediately.
16. Your next ED admission is a 65-year-old man who has shortness of breath as his main complaint, and the team wants to determine the probability of Pulmonary Embolism. You estimate that his Wells score is 3 points. Which one of the following statements is TRUE? Please select one.
- A. You should order a D-dimer test, which is highly sensitive, and proceed depending on the result.
 - B. Anticoagulation should be offered immediately without any further testing.
 - C. His risk of having PE is now estimated as high.
 - D. His probability of having PE is now estimated as near zero.
 - E. None of the above.
17. You are part of a team deciding whether to start using a moderately expensive test to screen for a disease. Which one of the following characteristics would make your team more likely to be in favor of screening? Please select one.
- A. The test has high sensitivity and specificity.
 - B. There is an efficacious treatment for the disease.
 - C. The natural history of the disease is one of high morbidity and mortality.
 - D. A, B, and C are true.
 - E. A and B are true, but not C.
18. The public health authorities in Guatemala wished to determine whether the use of mosquito nets reduces the risk of malaria. They identified 650 consecutive cases of malaria that were diagnosed over a period of 12 months. They matched them in a 1:3 ratio to people residing in the same towns, of the same gender, and of similar age (plus/minus 2 years). Researchers blinded to the participants' malaria diagnosis obtained a 20-question survey, which only 3 subjects failed to complete; the survey included several items inquiring about their use of mosquito nets, with particular emphasis on differentiating occasional use versus consistent daily use.

What study design did they use? Please select the one TRUE answer.

- A. Cross-sectional design
 - B. Consecutive case series
 - C. Case-control design
 - D. Prospective cohort design
 - E. Nested case-control design
19. Which one of the following is a potential concern for that study performed in Guatemala? Please select the one TRUE answer.
- A. Protopathic bias
 - B. Outcome ascertainment bias
 - C. Post-test referral bias
 - D. Recall bias
 - E. Prescription bias
20. When analyzing the association between lack of consistent daily use of mosquito nets and the risk of malaria in that study, which one of the following statistics would have been most appropriate? Please select the one TRUE answer.
- A. Log-rank test
 - B. Hazard ratio
 - C. Odds ratio
 - D. Cochran's Q
 - E. Relative risk
21. You are seeing a 65-year-old man at your clinic for a routine visit. He requests your advice about a new screening test for prostate cancer. He brings a brochure from another institution promoting the test, based on a recent study. That study randomized men older than 65 into two groups: one that was screened for prostate cancer using the new test and another that did not undergo any screening and in whom any cases of prostate cancer were diagnosed by their clinicians. The study found that median survival after diagnosis of prostate cancer was increased by 2 years in the screened group compared with the non-screened group. The brochure does not report whether the study found any differences in overall mortality or cancer-related mortality. In the group that was screened for prostate cancer, there were more cases of prostate cancer diagnosed overall, most of which were low-grade cancers. In the group that was not screened, there were significantly

fewer cases of prostate cancer diagnosed overall; however, those that were diagnosed had a more aggressive course.

Which of the following is a concern, based on what you know about this study? Please select one TRUE answer.

- A. Selection bias.
 - B. Contamination bias.
 - C. Observer bias.
 - D. Spectrum bias.
 - E. Length-time bias.
22. A Cardiology research network wants to confirm that the use of a new device has resulted in better outcomes after transcatheter aortic valve replacement. The older device had been the only one available for 4 years, then the new device was introduced and they were both used for 2 years. After that, only the new device has been used for the last 3 years. They have 12-month follow up data regarding outcomes for 91% of the patients, which was obtained by a team blinded to device type. They propose to compare outcomes corresponding to the initial 4 years and the final 3 years during which each device was exclusively used. What is the concern you should raise about their proposal? Please select one TRUE answer.
- A. Recall bias.
 - B. Ascertainment bias.
 - C. Non-adherence bias.
 - D. Interviewer bias.
 - E. Chronology bias.
23. You wish to compare two different prescription anti-inflammatory drugs, in regards to their known adverse effect: a higher risk of suffering a myocardial infarction. Preliminary evidence strongly suggests that the highest risk of myocardial infarction falls within a 1-month window after either drug is initiated. You plan to use all available data from a very complete database that prospectively collected information about the entire population of the Netherlands for the last 4 years. The database contains data about all prescriptions and fatal and non-fatal myocardial infarctions in that time period. The two drugs of interest have been in the Netherlands market for 6 and 3 years, respectively. You propose a model that will

capture incident cases of myocardial infarction in any patient using either drug during follow-up.

Which one of the following biases is a concern, given the current design? Please choose one TRUE answer.

- A. Protopathic bias.
- B. Length-time bias.
- C. Attrition of those susceptible bias.
- D. Recall bias.
- E. Lead-time bias.

24. Dapagliflozin has been shown for quite some time to improve renal outcomes in patients with diabetes in several randomized trials. You are concerned, however, that those trial results may not agree with your personal observations. You propose an observational study using the excellent database your hospital network has collected. You aim to identify all patients with a diagnosis of diabetes during the last 3 years, the period during which dapagliflozin has been in clinical use for that indication. Within those patients with diabetes you define two groups: one receiving dapagliflozin and the other not receiving dapagliflozin. You proceed to identify who progressed to end stage renal disease or advanced renal failure during the study period. You estimate the Relative Risk of progression.

Which one of the following could the Relative Risk estimate be compromised by? Please select one.

- A. Interviewer bias.
- B. Recall bias.
- C. Effect modification.
- D. Confounding by indication
- E. None of the above

25. You are a scientific member of the Institutional Review Board at your medical center. You are asked to review a protocol that the chief of Neurology is submitting, as the principal investigator. It is an open label, Phase 2 study to evaluate two different dosing strategies of a new drug to treat Huntington's disease. The drug is administered intrathecally. Phase 1 studies showed promising results reducing progression of Huntington's, but also the possibility of serious neurologic side effects.

Which of the following answers is FALSE? Please select one.

- A. This is a high risk/high yield proposal.
 - B. It is acceptable to use intrathecal injections to deliver the drug.
 - C. Complications and side effects must be actively monitored and periodically reported.
 - D. It is acceptable to obtain consent without first determining the person has capacity to provide consent.
 - E. The consent form should make unequivocally clear whether there are any institutional or personal conflicts of interest.
26. A study is proposed to assess a new type of drug eluting stent to treat coronary artery stenosis. This stent is expected to have a lower number of late re-stenosis events. It will be compared to the best standard of care, everolimus-eluting stents. The participants will be patients with multiple (two or more) significant lesions and have an indication of stenting due to chronic stable angina. Most of the patients will be referred from several other institutions to the interventional cardiology suites where the stenting will be performed. Those institutions serve different neighborhoods with ethnically and racially diverse populations. The design will mask the type of stent being used to the participants, to the follow-up staff, and to the cardiologists assessing whether the outcome takes place. Informed consent will be obtained on the day of the procedure, but participants will be contacted at least 1 week prior to that, in order to ensure that they receive an IRB-approved pamphlet explaining the study in advance. The pamphlet will include a phone number to call in case of questions. That phone number may also be used by the participant to decline participation in advance, without any repercussions on their clinically indicated procedure. All participants will be followed for at least one year. Please choose the FALSE statement from the list below.
- A. The informed consent process addresses adequately the challenges of same day consent.
 - B. Adequate consent respects the participants' autonomy.

- C. There are no obvious issues regarding the principle of Justice.
 - D. The control arm treatment complies with the principle of Beneficence.
 - E. It is inappropriate to enroll participants being referred from other centers.
27. Your laboratory has completed basic research about a new oral drug to treat COVID-19. You submit a protocol to your local IRB for a Phase 1 study of the drug. Which one of the following statements about the protocol is TRUE? Please choose one.
- A. Pregnant women may volunteer to participate.
 - B. Your laboratory team may volunteer to participate.
 - C. Residents of a nearby homeless shelter may volunteer to participate.
 - D. The drug may be administered to 10 participants per day from the start.
 - E. None of the above is correct.
28. The Autonomy principle in clinical research means (please choose one):
- A. Informed consent must be obtained and documented.
 - B. The participant may refuse further participation at any time.
 - C. The participant may demand their spouse be also enrolled as participant at any time.
 - D. A and B are correct.
 - E. B and C are correct.
29. The numbers at risk are found below the horizontal axis of Kaplan-Meier (KM) curve graphics. Which one of the following statements is TRUE about those numbers at risk?
- A. They represent the higher risk patients remaining in each study arm
 - B. They represent the effects of adverse effects in each study arm
 - C. They provide information about the precision of that portion of the KM curves.
 - D. They are inversely related to the power of our randomized sample.
 - E. They are equivalent to a period-adjusted number needed to treat.

30. A study of a new parenteral antibiotic for meningococcal meningitis confirmed the expectation of lower mortality rates for that antibiotic, as compared to the standard of care ($P=0.001$ for the difference). The Absolute Risk Difference was 0.14. Which one of the following statements about the results is TRUE?
- A. The standard of care is superior to the new treatment.
 - B. The Null Hypothesis regarding superiority of the new antibiotic cannot be rejected.
 - C. The new treatment is inferior to the standard of care.
 - D. The Number Needed to Treat was 7.
 - E. The Relative Risk was 7.
31. Your Oncology research team is planning a study of a new drug that treats advanced non-small cell lung cancer. This drug is expected to increase survival times and improve quality of life substantially. However, it is not expected to cure this very aggressive cancer. Which one of the following do you believe would be most informative to assess treatment benefits at the end of the trial? Please choose one answer.
- A. Chi-square.
 - B. Cox Proportional Hazards Model.
 - C. Wilcoxon Rank Sum
 - D. I-Squared
 - E. Relative Risk
32. A study of a new type of inhaler therapy for COPD shows a reduction in Hospital Admissions, as compared to the standard of care, over a median 24-month follow-up. The cumulative admission rate in the treatment arm was 14%, whereas the admission rate in the standard of care arm was 19% ($P=0.02$ for the difference). Which one of the following statements is TRUE? Please choose one.
- A. The Relative Risk Reduction was 0.05.
 - B. The Number Needed to Harm was 20.
 - C. The Absolute Risk Difference was 0.14.
 - D. The Number Needed to Treat was 20.
 - E. None of the above is correct.
33. A randomized clinical trial is proposed to assess a new type of drug eluting stent to treat coronary artery stenosis. This stent is expected to have a lower number of late re-stenosis

events, leading to better angina relief. It will be compared to the best standard of care, everolimus-eluting stents. The participants will be patients with unstable angina with at least one significant lesion at the time of their diagnostic angiogram. Informed consent will be obtained on the day of the procedure. Randomization will be performed in a 3:1 ratio, intervention to control. Participants will remain blinded to the type of stent being used in their case. The same cardiologist who performs the stenting will make their best effort to interview each participant at least twice over 12 months to assess their clinical status and determine whether their angina is better than before the stenting. Please choose the one TRUE answer.

- A. Ascertainment bias is a concern, given the methods.
 - B. The 3:1 randomization ratio biases the design in favor of the intervention arm.
 - C. The 3:1 randomization ratio allows obtaining more efficacy and safety data for the control arm than for the intervention arm.
 - D. Two follow-up interviews pose an unnecessary burden on participants.
 - E. Blinding participants to the type of stent they receive is unethical.
34. A study is proposed to confirm the expected superiority of a new type of implantable automated defibrillator for patients with advanced heart failure. This is the first phase 3 trial studying this device. It has an open label design that compares it to the best available defibrillator already in use. The outcome of interest is mortality over 12 months. An independent data and safety monitoring board is charged with ensuring the ethical completion of the trial. The board proposes a monitoring plan with 3 interim analyses, using symmetric Pocock boundaries. At each of those interim analyses a recommendation whether to stop or proceed will be given. Please select the one TRUE answer from the list below.
- A. A non-inferiority design would have been more ethically appropriate.
 - B. Asymmetric boundaries, with a triangular design, would have been more ethically appropriate.

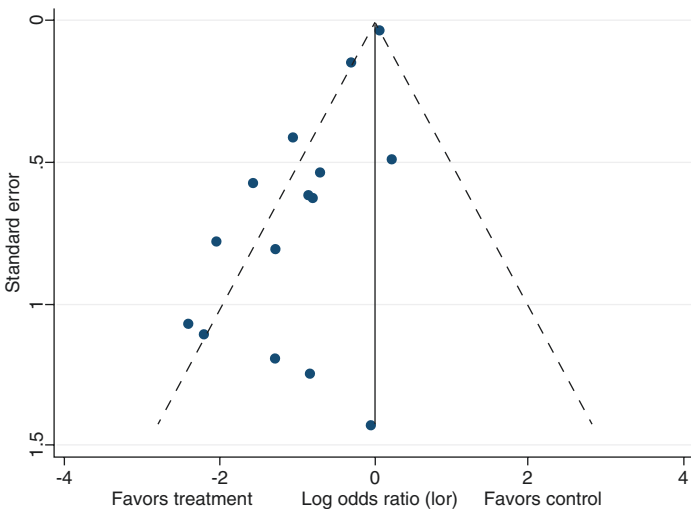
- C. A placebo-controlled design would have been more ethically appropriate.
 - D. The interim analyses increase the probability of type 2 statistical error.
 - E. The interim analyses increase the probability of confounding by indication.
35. Please select the one FALSE answer from the list below. In order to preserve the benefits of randomization we always try to do the following when we conduct a randomized clinical trial:
- A. Double-blinding of intervention.
 - B. Complete follow-up of both study arms.
 - C. Masked ascertainment of outcomes.
 - D. Timed cross-overs from one arm to another.
 - E. Unmasked interim analyses are performed by an independent safety monitoring board.
36. You are reading an article that reports that a new medication, delivered as a percutaneous patch, improves dyspnea symptoms in patients with advanced COPD. The study was randomized double-blinded, and provided patients in the control arm with an identical patch that contained no active substance. The two study arms had balanced distribution of potential covariates at baseline. Symptoms were assessed through monthly questionnaires administered by researchers who were blinded to randomization status. The consent form appropriately warned participants that the new medication could cause palpitations in 33% of the patients receiving it. Follow-up data was obtained in 97% and 96% of the intervention and control arms, respectively. Intervention fidelity was 98% in both arms. Outcomes were analyzed based on the intention to treat principle.
- Which one of the following statements about the study is TRUE at the time of analysis? Please choose one.
- A. There was risk of confounding by known covariates.
 - B. There was risk of interviewer bias.
 - C. There was risk of involuntary unmasking bias.
 - D. There was risk of cross-over bias.

E. There was risk of chronology bias.

37. A randomized clinical trial for a new treatment for septic shock is being planned. The initial sample size calculation was carried out to achieve 90% power to detect an estimated 20% reduction in mortality at a significance threshold of $P\text{-Value} < 0.05$. The cumulative mortality by the end of the trial was estimated as 33% for the standard of care arm.

Which one of the following changes would allow the researchers to perform the study with a SMALLER sample size? Please select one TRUE answer.

- A. An estimated cumulative event rate of 23% in the standard care arm.
 - B. An estimated 25% reduction in mortality.
 - C. A significance threshold of $P < 0.01$.
 - D. Power of 95% percent.
 - E. C and D are correct.
38. A meta-analysis is performed to assess the efficacy of Selective Serotonin Reuptake Inhibitors to treat depression in adolescents. The funnel plot shown below is generated. What does it strongly suggest? Please select one TRUE answer.



- A. There is verification bias.
 - B. There is post-test referral bias.
 - C. There is publication bias.
 - D. There is prescription bias.
 - E. There is protopathic bias.
39. You are asked to review a study. It is a double-blinded placebo-controlled phase 3 study to assess a new drug to treat advanced heart failure. The control arm includes all drugs that conform the standard of care, and placebo. The treatment arm also includes the standard of care and adds the new drug. Given the high mortality in this condition the outcome of interest is death. The new drug is expected to reduce mortality, although it will be more expensive, and probably will cause more side effects in a substantial proportion of patients. The investigators propose a non-inferiority design. Please select one FALSE answer.
- A. The outcome selected is relevant from a societal perspective.
 - B. A non-inferiority design, given the circumstances, is probably unethical.
 - C. It is acceptable to proceed immediately with the study design proposed by the investigators.
 - D. A superiority design, with predefined stopping boundaries for efficacy or unexpected harm, should be required.
 - E. The double-blinded design reduces the risk of bias.
40. Please select the one TRUE answer. The null hypothesis (H_0) in a non-inferiority clinical trial states that:
- A. The new treatment is superior to the standard of care.
 - B. The new treatment is inferior to the standard of care.
 - C. The new treatment is equivalent to the standard of care.
 - D. The new treatment is not inferior to the standard of care.
 - E. The standard of care is inferior to the new treatment.
41. A medical device manufacturer has created a new device for the intravascular closure of brain aneurysms. They approach a network of academic interventional radiologists with a proposal for a non-inferiority clinical trial. They state there is a reasonable expectation that their device will have similar efficacy with a lower risk of adverse effects, when compared to

the standard of care. The expected event rate for the standard of care is 33%. The study would have a non-inferiority margin of 8 percentage points. Which one of the following statements is TRUE? Please choose one.

- A. The non-inferiority margin is equivalent to a relative risk margin of 1.24.
- B. The non-inferiority margin is equivalent to a number needed to harm of 12.
- C. The non-inferiority margin for absolute risk increments.
- D. The non-inferiority margin for relative risk increments is 0.08.
- E. A, B, and C are correct.

42. A double-blinded randomized clinical trial was performed in a Brazilian hospital network to assess the efficacy of Paxlovid treatment in patients with COVID-19 infection and at least one risk factor for disease progression. At the end of the study the effect on the risk of death was reported as Relative Risk = 0.88; 95% Bayesian credible interval, 0.76 to 1.01. The estimated probability of the Relative Risk being <1 was 97%.

Which one of the following statements is FALSE? Please select one.

- A. The findings cannot be interpreted because the credible interval includes the value of $RR=1$.
 - B. The estimated Relative Risk Reduction was 12%.
 - C. The credible interval is the Bayesian equivalent of a confidence interval.
 - D. There is an estimated 95% probability that the population's Relative Risk lies between 0.76 and 1.01.
 - E. The probability of Paxlovid not reducing mortality risk was estimated as 3%.
43. In the same ivermectin study the investigators reported their statistical analysis as follows: "Posterior probability for the efficacy of Paxlovid with regard to the outcomes was calculated with the use of the beta-binomial model for the percentages of patients with an event, starting with uniform prior distributions for the percentages."

- Which one of the following statements is FALSE? Please select one.
- A. They used an uninformative prior probability.
 - B. Their method was likely to provide credible intervals that were very similar to frequentist confidence intervals.
 - C. The posterior probability is the Bayesian estimate of treatment benefit from ivermectin.
 - D. The posterior probability estimates did not depend on the study’s mortality results.
 - E. The Bayesian method allowed estimating probabilities of possible benefit or harm.
44. A meta-analysis comparing a new type of non-opioid oral analgesics to acetaminophen for chronic lower back pain is published. The Cochrane risk of bias graphic is shown below (each individual study is represented by a column). Please select the one TRUE answer.

																		Random sequence generation
																		Allocation Concealment
																		Blinding of participants and personnel
																		Blinding of outcome ascertainment
																		Incomplete outcome data
																		Selective reporting
																		Other bias

- A. There is serious concern about performance bias.
- B. There is serious concern about incomplete follow-up.
- C. There is serious concern about inappropriate randomization.
- D. There is serious concern about publication bias.
- E. There is serious concern about excessive heterogeneity.

45. That same meta-analysis provides the following statement in the Results section. "The efficacy of the new drug to reduce the need for opioids to treat episodes of back pain was higher than that of acetaminophen: Hazard Ratio (95% confidence interval) 0.84 (0.71-0.98); the I^2 statistic was 91%." What is your assessment of those results?
- A. Prescription bias should be addressed.
 - B. Recall bias should be addressed.
 - C. Selection bias should be addressed.
 - D. Publication bias should be addressed.
 - E. Excessive heterogeneity should be addressed.
46. You perform a study of ultra-fast CT scanning with intravenous angiography to diagnose Coronary Artery Disease (CAD). You prospectively recruit 650 consecutive patients who are suspected of suffering from angina pectoris. They all undergo both the CT test and an intracoronary angiogram. When you analyze the results, the proportion of all the patients with an abnormal CT angiography who had CAD is known as (please select one TRUE answer):
- A. Prevalence or pre-test probability.
 - B. Negative predictive value.
 - C. Positive predictive value.
 - D. Sensitivity.
 - E. Specificity.
47. In that same study the proportion of all patients without CAD who had a negative CT test is known as (please select one TRUE answer):
- A. Prevalence or pre-test probability.
 - B. Negative predictive value.
 - C. Positive predictive value.
 - D. Sensitivity.
 - E. Specificity.
48. Please select the one FALSE answer. A non-inferiority design is acceptable for a clinical trial when the new treatment is expected to have:
- A. Similar efficacy as the standard of care
 - B. Better tolerability

- C. Better acceptability
 - D. Similar efficacy as placebo
 - E. Lower cost
49. Which one of the following statements about the Bayesian analysis of clinical trials is TRUE? Please select one.
- A. An uninformative prior should usually be the main analytic approach
 - B. An extremely optimistic prior should usually be the main analytic approach.
 - C. A pessimistic prior should usually be the main analytic approach
 - D. Any type of prior probability could be used, as this is irrelevant.
 - E. A strict P-Value interpretation should be the main analytic approach.
50. Your research team is planning a study to compare the two most commonly used thrombectomy devices for acute stroke treatment. Preliminary data strongly suggests that active smokers are at higher mortality risk after any type of stroke, but they may benefit greatly from a thrombectomy. Their budget will not allow a large sample size. Which one of the following statements is TRUE regarding the study they are planning? Please select one.
- A. Excluding smokers would increase their external validity.
 - B. They should consider stratified randomization.
 - C. They should change their design to a non-randomized observational study in non-smokers.
 - D. Possible confounding is not a concern.
 - E. To include active smokers would not be ethically appropriate.

Answers to Practice Questions

1C; 2D; 3D; 4D; 5B; 6A; 7B; 8B; 9A; 10E; 11C; 12A; 13 A; 14B;
15D; 16A; 17D; 18C; 19D; 20C; 21E; 22E; 23C; 24D; 25D; 26E;
27E; 28D; 29C; 30D; 31B; 32D; 33A; 34B; 35D; 36C; 37B; 38C;
39C; 40B; 41E; 42A; 43D; 44A; 45E; 46C; 47E; 48D; 49A; 50B.

Open Access Evidence Based Clinical Knowledge

The Cochrane Collaboration: If you only visit one EBM site ever, this should be it. The Cochrane Library includes Systematic Reviews and Meta-Analysis, clinical answers based on evidence, and a registry of clinical trials.

<https://www.cochrane.org/>

UK NATIONAL INSTITUTE FOR HEALTH AND CARE RESOURCES: Open the Clinical Knowledge Summaries tab, to find evidence reviews for the most relevant clinical topics.

<https://www.nice.org.uk/>

The Centre for Evidence-Based Medicine: This site provides essential, up-to-date, evidence based reviews. It has promptly provided links for COVID-19 related data.

[Home - 2020 - The Centre for Evidence-Based Medicine \(cebm.net\)](#)

Oxford University: Oxford University has been a leader in the evidence-based approach medicine. That has included the fight against COVID-19. Their site gives us access to their work, which provides a template for a thoughtful response to a pandemic.

<https://www.research.ox.ac.uk/area/coronavirus-research>

BANDOLIER: The Knowledge Library link gives access to itemized evidence reviews in clinically relevant topics, that are very useful.

<http://www.bandolier.org.uk/>

NIHR Evidence: Clinically oriented topic reviews. From the UK National Institute for Health and Care Research

<https://evidence.nihr.ac.uk/>

PubMed Clinical Queries: This PubMed tool allows the use of filters to search for evidence regarding therapy, clinical prediction guides, diagnosis, etiology, or prognosis.

<https://pubmed.ncbi.nlm.nih.gov/clinical/>

NNT.com: This site starts off with a clear description of Numbers Needed to Treat. Tabs take you to NNT curated by specialty.

<https://www.thennt.com>

Online EBM Resources That Require Institutional Access

JAMA EVIDENCE: If your institution has access to this site, you will find the outstanding series of foundational EBM article series, as well as clinical calculators, and the JAMA Guide to Statistics and Methods

<https://jamaevidence.mhmedical.com>

BMJ Statistics Notes: Each article provides a user-friendly introduction to relevant statistical issues.

[Statistics notes | The BMJ](#)

BMJ Best Practices: targeted to either clinicians or institutions, this site provides updated guidelines for all clinical fields, and includes access through a mobile app.

<https://bestpractice.bmj.com>

EBM Calculators, Decision Making Tools

EBMcalc:

<https://ebmcalc.com>

MDcalc:

<https://www.mdcalc.com>